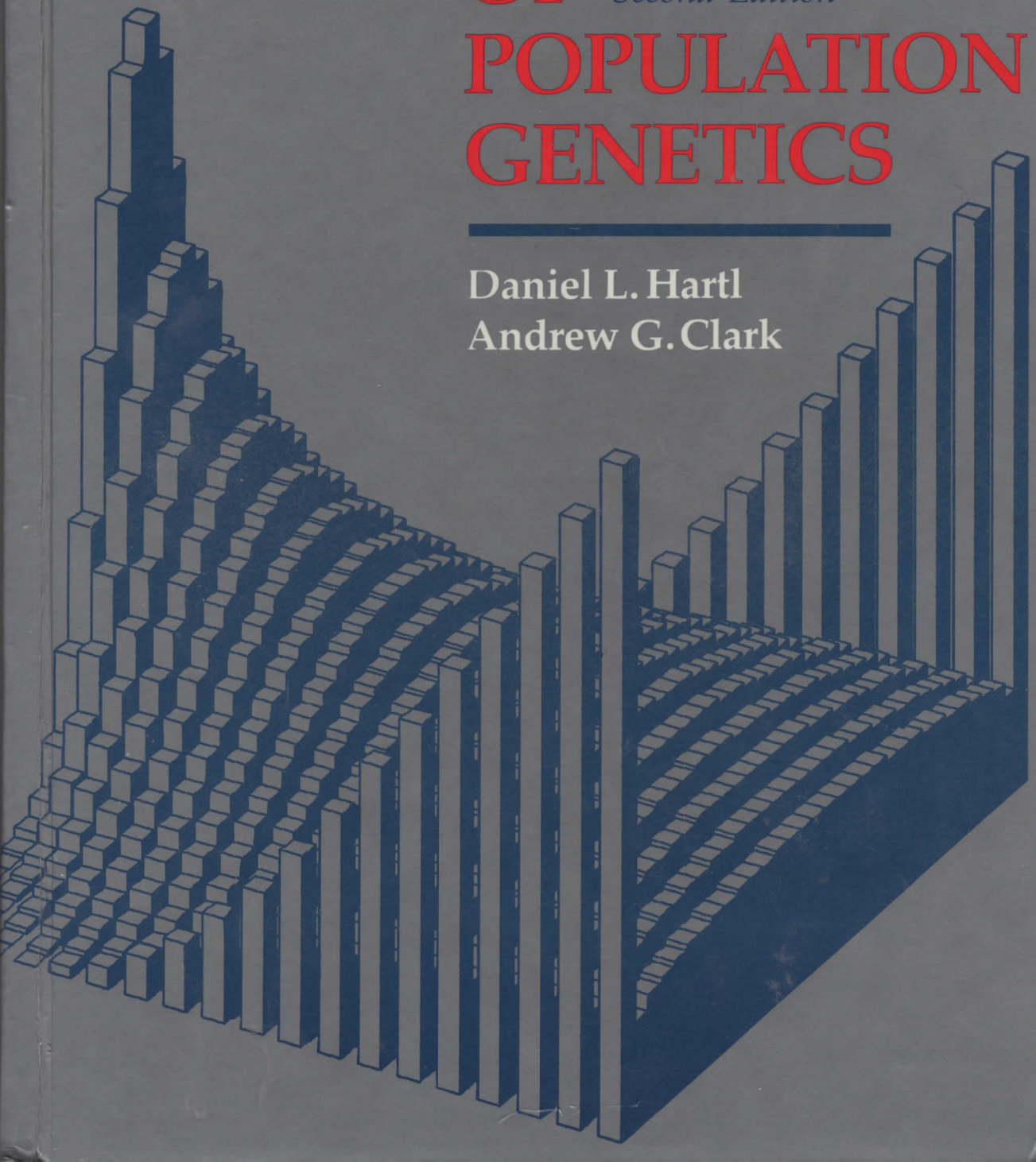
PRINCIPLES OF POPULATION GENETICS

Second Edition

Daniel L. Hartl
Andrew G. Clark



Robert V Miller
Maywood
13 Feb, 1989

Principles of Population Genetics

DANIEL L. HARTL
WASHINGTON UNIVERSITY SCHOOL OF MEDICINE

ANDREW G. CLARK
PENNSYLVANIA STATE UNIVERSITY

PRINCIPLES OF POPULATION GENETICS

SECOND EDITION

SINAUER ASSOCIATES, INC. • PUBLISHERS
Sunderland, Massachusetts

To Christopher

PRINCIPLES OF POPULATION GENETICS, Second Edition

Copyright © 1989 by Sinauer Associates, Inc. All rights reserved. This book may not be reproduced in whole or in part for any purpose whatever, without permission from the publisher. For information address Sinauer Associates, Inc., Sunderland, Mass., 01375 U.S.A.

Library of Congress Cataloging-in-Publication Data

Hartl, Daniel L.
Principles of population genetics.

Bibliography: p.
Includes index.
1. Population genetics. 2. Quantitative genetics.
I. Clark, Andrew G., 1954– II. Title.
QH455.H37 1988 575.1'5 88-30818
ISBN 0-87893-302-6

Printed in U.S.A.

5 4 3 2 1

Contents

Preface xiii

CHAPTER ONE

Darwinian Evolution in Mendelian Populations 1

The Scope of Population Genetics 1
Mathematical Models in Population Genetics 3
The Empirical Problem 4
Diversity of Populations 6
Phenotypic Variation 6
Simple Genetic Traits 7
Box A The Maxam-Gilbert Method of DNA Sequencing 15
Allele Frequencies and Polymorphism 17
Box B Calculating Allele Frequencies 18
Allozyme Polymorphisms 19
Box C Allozyme Data: A Simple Example 21
Molecular Sequence Variation 25
Box D Restriction Site Variation 25
Box E Allelic Variation in DNA Sequence 30
The Hardy-Weinberg Principle 31
Random Mating and Random Union of Gametes 31
One Locus with Two Alleles 32
Box F Random Mating and Random Union of Gametes 34
Box G De Finetti Diagram for One Locus and Two Alleles 35
The Hardy-Weinberg Principle in Operation 37
The Hardy-Weinberg Principle with Three or More Alleles 39
Box H Estimating Allele Frequencies Using the EM Algorithm 40
The Hardy-Weinberg Principle for X-linked Genes 43
Box I Hardy-Weinberg Principle for X-linked Genes 44
The Hardy-Weinberg Principle for Two Genes 45
Box J Linkage Disequilibrium and the Correlation Coefficient 53
Estimating Linkage Disequilibrium 55
Box K Estimating Linkage Disequilibrium for Two Diallelic Genes 56
Problems 57

CHAPTER TWO

Random Genetic Drift 61

Binomial Sampling 62
The Wright-Fisher Model of Random Genetic Drift 66
Box A Markov Chains and the Wright-Fisher Model 67
The Diffusion Approximation 71
Parallels Between Random Drift and Inbreeding 74
Effective Population Size 82
Fluctuation in Population Size 83
Box B Effective Population Size 85
Unequal Numbers of Males and Females 86

Overlapping Generations 87
Non-Poisson Distribution of Fecundity 88
Geographic Dispersion of Populations 90
Genealogy and the Coalescent, Or, All About Eve 90
Problems 93

CHAPTER THREE

Mutation and the Neutral Theory 96

Origin of Genetic Variation 96
Classes of Mutation 97
Estimation of Mutation Rate 103
Spontaneous Lethal Mutation Rate in *Drosophila* 103
Coat Color Mutations in Mice 107
Mutation Rate of Allozymes in *Drosophila* 108
Mutation Rate of Polypeptides in Humans 108
Is the Mutation Rate Adaptive? 109
Rate of Change of Allele Frequency Under Mutation Pressure 110
Box A The Dynamics of Mutation 112
Box B Bidirectional Mutation 114
Is Mutation Random at the Molecular Level? 114
Effects of Spontaneous Mutations on Fitness 118
Infinite Alleles Model 122
Balance Between Mutation and Random Genetic Drift 122
Box C Equilibrium Autozygosity for the Infinite Alleles Model 124
Box D Effective Number of Alleles 126
The Ewens Sampling Formula 126
Two Neutral Genes and the Infinite Alleles Model 130
Box E Correlation of Heterozygosities of Linked Neutral Genes 132
Infinite Sites Model 132
The Coalescent under the Infinite Sites Model 134
Charge State Model 135
Tests of Selective Neutrality 136
Box F The Ewens-Watterson Test 141
Problems 143

CHAPTER FOUR

Natural Selection 147

Definitions of Fitness 148
Complications in Measuring Fitness 149
Classical Models (Viability) 151
One Gene with Two Alleles 151
Box A Numerical Example of Allele Frequency Change with Selection 153
Box B Analysis of Stability of Equilibria 159
The Fundamental Theorem of Natural Selection 164
Multiple Alleles 168
The Fundamental Theorem with Multiple Alleles 176
Two-Sex Viability Selection 177
X-Linked Viability Selection 180
Haploid Selection 182
Box C Observing Natural Selection in a Chemostat 183
Frequency-Dependent Selection 184
Fecundity Selection 186
Meiotic Drive and Gametic Selection 190
Box D Meitoic Drive of the t *Alleles in* Mus 192
Sexual Selection 194
Box E Sexual Selection in Monogamous Birds 197
Balance of Opposing Forces 198
Mutation–Selection Balance 199
Meiotic Drive and Viability Selection 201
Viability Selection in Haploid and Diploid Phases (Selection in Sporophytes and Gametophytes) 204
Viability and Fertility Selection 206
Viability Selection in Finite Populations 206
More Than One Gene 209
Two-Locus Selection and Epistasis 209
Box F Evolution of the Rate of Recombination 216

Multilocus Evolution 217
Multilocus Linkage Disequilibrium 219
Selection Component Analysis 220
Estimation of Selection Components in *Drosophila* 220
Box G Estimation of Selection Components in Experimental Populations of Drosophila melanogaster 222
Estimation of Fitness Effects Using Selective Media 223
Estimating Components of Selection Using Samples from Natural Populations 224
Molecular Variation and Natural Selection 228
Ecological Complications 230
Problems 232

CHAPTER FIVE

Inbreeding and Other Forms of Nonrandom Mating 235

The Inbreeding Coefficient 236
Calculation of the Inbreeding Coefficient from Pedigrees 238
Box A Inbreeding in Irregular Pedigrees 241
Box B Inbreeding Coefficient for X-Linked Genes 243
Genotype Frequencies with Inbreeding 244
Box C F *is the Correlation of Uniting Gametes* 249
Effects of Inbreeding and the Frequency of Human Genetic Disorders 250
Regular Systems Of Inbreeding 255
Box D Inbreeding Coefficient with Recurrent Sib Mating 258
The Mixed Mating Model (Partial Selfing) 261
Box E One-Locus Genotype Frequencies Under the Mixed Mating Model 263
Inbreeding and Two Loci 263
Selection and Inbreeding 268
Assortative Mating 271
Positive Assortative Mating 271
Negative Assortative Mating 274
Problems 278

CHAPTER SIX

Population Subdivision and Migration 281

Wahlund's Principle 282
The Wahlund Effect with One Gene with Two Alleles 282
Box A The Wahlund Effect in Scandinavian Cod 285
The Wahlund Effect with One Gene and Many Alleles 288
The Wahlund Effect with Two Genes 291
Wright's *F* Statistics 293
Box B The Wahlund Effect Assures that $H_T > \overline{H}_S$ 294
Box C Demonstration that F_{IS}, F_{ST}, *and* F_{IT} *are Related by the Equation* $(1 - F_{IS})(1 - F_{ST}) = (1 - F_{IT})$ 296
Box D Estimating F *Statistics from Population Samples* 297
Box E Cockerham's Partitioning of F *by Analysis of Variance* 298
Box F Population Structure of Yanomama, Gainj, and Kalam Peoples 300
Race and Genetic Distance 301
Box G Genetic Distance 303
The Island Model of Migration 306
One-Way Migration 308
One-Way Migration and Selection 309
Migration in Finite Populations (Wright's Island Model) 310
The Island Model with Mutation 314
Box H Island Model for Mitochondria 316
Stepping Stone Models 317
Estimating Migration Rates 319
Interdeme Selection and the Shifting Balance Theory 323
Box I Human Isolates and the Shifting Balance Theory 324
Problems 326

CHAPTER SEVEN

Molecular Population Genetics 329

The Tools 330
Southern Blotting and Inferences from Restriction Site Variation 331
Box A Restriction Map Variation in the Adh *Gene of* Drosophila melanogaster 335
DNA Sequencing 339
Box B Sanger Dideoxy Chain-Termination Method for Sequencing DNA 341
Box C The Polymerase Chain-Reaction Method of DNA Amplification 343
DNA–DNA Hybridization 344
Box D DNA Sequence Alignment Methods 345
Microcomplement Fixation 347
The Neutral Theory and Molecular Evolution 349
Theoretical Principles of the Neutral Theory 350
Rates of Amino Acid Replacement 354
Rates of Nucleotide Substitution 357
Molecular Clocks 358
Box E Calculating the Average Rate of Molecular Evolution 361
Box F Is the Molecular Clock Episodic? 364
Patterns of Nucleotide and Amino Acid Substitution 368
Box G Calculating the Number of Synonymous and Nonsynonymous Sites 370
Inference of Natural Selection from Nucleotide Sequence Data 376
Phylogenetic Inference and Gene Genealogies 377
Box H Inference of Recombination from Nucleotide Sequence Data 382
Evolution in Multigene Families 384
Mechanisms of Concerted Evolution 384
Duplication and Divergence in Multigene Families 388
Box I Estimating the Time of a Duplication 390
Ribosomal RNA Gene Family 393
Cryptic Genes 395
Multigene Superfamilies 396
Localized Highly Repetitive DNA Sequences 399
Dispersed Highly Repetitive DNA Sequences 402
Box J Rate of Disappearance of Repetitive DNA Sequences 405
Transposable Elements 407
Factors Controlling the Population Dynamics of Transposable Elements 407
Insertion Sequences and Composite Transposons in Bacteria 409
Box K Examples Where an Insertion Sequence and a Tn *Element Appear to be Selectively Favored* 411
Transposable Elements in Eukaryotes 413
Hybrid Dysgenesis in *Drosophila* 415
Endogenous Retroviruses in Mammals 416
Box L Calibrating DNA–DNA Hybridization with the Molecular Clock 418
Mitochondrial and Chloroplast DNA Evolution 421
Box M Human mtDNA Divergence 423
Population Genetic Theory for Cytoplasmically Transmitted Elements 425
Problems 427

CHAPTER EIGHT

Evolutionary Genetics of Quantitative Characters 430

Types of Quantitative Traits 431
Box A The Normal Distribution and the Central Limit Theorem 433
Resemblance between Relatives and the Concept of Heritability 435
Box B Statistical Tools Used in Quantitative Genetics 436
Box C Calculation of Heritability from Parent-Offspring Regression 440
Artificial Selection and Realized Heritability 441
Prediction Equation for Individual Selection 442

Box D Realized Heritability of Bristle Number in Drosophila 445
Selection Limits 446
Genetic Basis of Quantitative Traits 450
Change in Gene Frequency 454
Genetic Model for the Change in Mean Phenotype 456
Components of Phenotypic Variance 458
Genetic and Environmental Sources of Variation 458
Components of Genotypic Variation 463
Box E Heritability of Simple Mendelian Traits 465
Covariance between Relatives 466
Norm of Reaction and Phenotypic Plasticity 472
Box F Norm of Reaction 473
Experimental Assessment of Genetic Variance Components 476
Box G Components of Genetic Variance in Natural Populations 477
Twin Studies and Inferences of Heritability in Humans 480
Threshold Traits and the Genetics of Liability 481
Estimation of the Number of Genes Affecting a Quantitative Character 485
Box H Estimating the Number of Genes Determining Corn Oil Content 488
Correlated Response and Genetic Correlation 489
Box I Genetic Correlations among Enzyme Activities 490
Box J Correlated Response to Selection on Weight Gain in Rats 492
Selection on Multivariate Characters 493
Multivariate Selection Theory 493
Box K The Multivariate Gaussian Distribution 494
Inference of Selection from Phenotypic Data 496
Random Genetic Drift and Phenotypic Evolution 498
Evolution of Quantitative Traits: Mutation-Selection Balance 501
Mendelian Markers and Quantitative Traits 504
Mapping Genes that Influence Quantitative Characters 505
Selection for Continuous Traits and Response of Identified Genes 507
Problems 511

CHAPTER NINE

Ecological Genetics and Speciation 514

Density-Dependent Natural Selection 517
Box A Experimental Tests of Density-Dependent Natural Selection 522
Natural Selection in Heterogeneous Environments 524
Box B Temporal Variation in Selection Coefficients 529
Age-Structured Populations and the Evolution of Senescence 530
Demographic Theory 531
Selection in Age-Structured Populations 533
The Evolution of Senescence 535
Box C Experimental Studies of Senescence in Drosophila 537
Physiological Genetics, Allozymes, and Just-So Stories 538
Allozymes and Metabolic Control 539
Box D Population Genetics of Phenylketonuria 543
Ecological Genetics of Single-Gene Substitutions 543
Heavy Metal Tolerance in Plants 548
Box E Ecological Genetics of Mycophagous Drosophila 549
Heterozygosity and Growth Rate 553
Sex Ratio 555
Group Selection, Kin Selection, and the Evolution of Cooperation 560
Group Selection 561
Kin Selection and the Evolution of Altruism 563

Box F Kin Selection in Haplodiploid Insects 566
Evolution of Cooperation 566
Coevolution 569
Mimicry 570
Industrial Melanism in Moths 577
The Evolution of Antibiotic Resistance in Bacteria 580
Coevolution of Bacteria and Bacteriophage 586
Speciation 586
Box G The Species Concept 588
Modes of Reproductive Isolation 589
Geographical Aspects of Speciation 590
Genetics of Species Differences 597
Box H Genetics of Hybrid Sterility in Drosophila 600
Box I The Bizarre Genetics of Poeciliopsis 603
Problems 604

Suggestions for Further Reading 607

Answers to Chapter-End Problems 610

Bibliography 628

Author Index 665

Subject Index 670

Preface

This is an exciting time for population genetics because it is becoming united with molecular biology. Application of molecular methods to evolutionary studies has already illuminated many aspects of evolutionary history. Population genetics is essential in interpreting the molecular sequences of genes and proteins and in using sequences to make inferences about evolutionary history. Population genetics is also concerned with the processes and mechanisms by which evolutionary changes occur. Being one of the most quantitative branches of biology, population genetics makes extensive use of mathematical and statistical concepts to describe and understand populations of organisms. While experimental population genetics has been invigorated by the application of molecular techniques, theoretical population genetics has also undergone a resurgence because the molecular data make it possible to test many of the predictions of the theory. Population geneticists generally have broad interests because the field cuts across so many disciplines—molecular biology, genetics, ecology, evolutionary biology, systematics, natural history, plant breeding, animal breeding, conservation and wildlife management, human genetics, sociology, anthropology, mathematics, and statistics. This book is our attempt to convey the excitement and enthusiasm shared by our colleagues in the field.

The second edition of *Principles of Population Genetics* is intended for third and fourth year undergraduates as well as graduate students. In addition to emphasizing traditional aspects of the field, it includes most of the new concepts and types of data emerging from molecular methods. No mathematical background beyond elementary calculus is needed, although we have included verbal discussions of models that require more advanced methods for their analysis. Special pains have been taken to motivate and justify the theoretical models and to put them in the broader context. Many concepts are illustrated by numerical examples, using actual data wherever possible, and special topics and examples are often set off in boxes. Each chapter ends with about 20 problems; their solutions are worked out in full at the end of the text.

The book builds sequentially through concepts of genetic variation and its measurement, various forces that influence genetic variation, molecular pop-

ulation genetics, quantitative genetics, and ecological genetics. Chapter 1 deals with surveying the types of genetic variation in natural populations and incorporates the results of molecular methods ranging from from restriction mapping to DNA sequencing. Chapter 2 deals with random genetic drift — the change in allele and genotype frequencies resulting from random sampling from one generation to the next in finite populations.

Chapters 3 and 4 consider experimental and theoretical aspects of mutation and natural selection. The process of mutation is now known to be more complex and less random than was originally imagined, and the consequences of this are discussed. Natural selection is the force that Darwin saw as responsible for adaptive evolution. In our treatment, we focus on selective processes that can maintain genetic variation. These chapters also introduce the neutral theory of molecular evolution — not as an alternative to classical Darwinian evolution, but as a mechanism governing the fate of the sort of molecular variation that does not have significant direct effects on fitness.

Chapter 5 deals with inbreeding and mating systems, with emphasis on the concept of identity by descent. Classical models of inbreeding are given new life using graphical displays, and real examples illustrate many of the phenomena. Chapter 6 concerns migration and population structure and the importance of population subdivision in fitting genetic models to data.

Chapter 7 covers the vast array of rapidly accumulating data on molecular evolutionary genetics. A theme that unifies the study of molecular evolution is the neutral theory, and a close examination is made of the correspondence between this theory and relevant data. The phenomenon of the molecular clock of nucleotide substitutions and amino acid replacements is discussed in view of recent findings that the variance in substitution rate is greater than expected under the simplest version of the neutral mutation model. Forces at the molecular level can direct evolutionary change in ways that are very different from those of classical population genetics. These include concerted evolution in multigene families, transposable elements, and processes affecting organelle genomes.

Quantitative genetics is the subject of Chapter 8. Because adaptive evolution takes place at the level of the phenotype, quantitative genetics provides the tools for understanding phenotypic evolution. We emphasize the poor correlation that exists between evolution at the phenotypic and molecular levels.

Chapter 9 covers a range of selected topics in ecological genetics. These include density-dependent selection, selection in heterogeneous environments, age-structured populations and the evolution of senescence, physiological genetics, heterozygosity and growth rate, kin and group selection, sex ratio, and coevolution.

The real purpose of the book is to excite and challenge students because,

more than an assemblage of facts and formulas, population genetics is a way of thinking — thinking at the population level. The field is dynamic and changing. It does not consist of a list of dry equations covered in dust. For this reason we have given attention to the framework of population genetics so students can develop a feel for the models and what they mean, and have emphasized that many questions remain unresolved and in need of careful and imaginative work.

ACKNOWLEDGMENTS

This book was greatly improved by the efforts of many people. The entire text of the first edition was encoded into computer files by Kay Webb. Thorough chapter reviews of the new edition were provided by Chip Aquadro (1), Larry Mueller (2), Tom Whittam (3 and 6), Jim Curtsinger (4), Bruce Weir (5), Cathy Laurie (7), Michael Turelli (8), and Rollin Richmond (9). The time and effort that these reviewers put into the task was truly gratifying, and demonstrated the importance they place on good teaching tools. The staff at Sinauer Associates did a splendid job of easing us through the revision. Andy Sinauer shared our trials and successes and provided an amiable and supportive atmosphere. Steve Miller made many improvements in the manuscript in his job as copy editor, and patiently tracked down elusive references. Joe Vesely proved himself fussier than us in producing clear graphics and a crisp, clean design. Carol Wigg spent many nights dreaming about superscripts and subscripts as she coded the files for electronic typesetting. Barbara Andersen got stuck with a daunting crossword puzzle when she agreed to do the index. In Clark's lab, Sharyn Houtz, Lisa Keith, Eva Lyckegaard, Susan Sweeney, and Francis Szumski all contributed to the book by tracking down references, proofreading chapters, commenting on esthetic merits of figures, and keeping the lab running smoothly. The same essential tasks were carried out in the Hartl lab by James Ajioka, Glenn Bryan, John Carulli, Robert DuBose, Danny Garza, Ann Gerber, Robert Jones, Jong-Bong Kim, Jeffrey Lawrence, Ann MacPeek, Kyoko Maruyama, Meetha Medhora, Howard Ochman, Tina Rouse, Ann Vellek, and Kay Webb.

DANIEL L. HARTL

ANDREW G. CLARK

Principles of Population Genetics

Chapter One

Darwinian Evolution in Mendelian Populations

THE SCOPE OF POPULATION GENETICS

Population genetics is the study of how Mendel's laws and other genetic principles apply to entire populations. Such a study is essential to a proper understanding of evolution because, fundamentally, evolution is the result of progressive change in the genetic composition of a population. Population geneticists seek to understand and predict the effects on populations of such genetic phenomena as segregation, recombination, transposition, and mutation. At the same time, population geneticists take into account such ecological and evolutionary factors as population size, patterns of mating, geographic distribution of individuals, migration, and natural selection. The many genetic, ecological, and evolutionary factors that influence populations also interact, one upon the other. To gain an understanding of such complex interactions is a formidable task, requiring a variety of approaches including purely theoretical model building, careful descriptive observations, and interpretation of statistical patterns, as well as controlled manipulative experiments.

Ideally, one would wish to know how to describe the types and frequencies of genes in a population, to explain how the population's genetic composition came to be the way it is in order to predict how the population would change as a result of natural selection or as a result of artificial selection applied by a plant or animal breeder. Population geneticists are a long way from being able to do these things as well as they might be done, but in the 90 or so years since the formal study of population genetics was inaugurated, considerable progress has been made. Indeed, a balanced view of the recent history

of biology would not permit the brilliant and relatively rapid success of molecular biologists to completely overshadow the slow, steady progress of population geneticists against their own problems, problems that can be particularly stubborn because they are often so subtle and sometimes maddeningly complex.

Remarkable progress has been made over the last ten years in characterizing molecular evolution. However, it is important to consider the extent to which molecular methods are limited when applied to solving the problems of population genetics. Questions concerning the structure of genetic variation and its partitioning among populations can be solved to a high degree of precision, but to the extent that population genetics is concerned with the genetic basis of evolutionary adaptation, molecular methods have not yet provided all of the answers. The term *adaptation* refers to the integration of the phenotype into its environment, and molecular techniques are still far from giving us a complete understanding of the processes of development that connect or "map" genotypes into phenotypes. It is precisely because of the central importance of this connection that scientists who regard themselves as population geneticists are hard at work in the laboratory trying to understand the molecular basis of such phenomena as gene regulation.

Despite these caveats, it is appropriate to point out that certain central problems in population genetics have been solved by use of relatively straightforward approaches and have yielded relatively simple principles; these principles have applications and implications for topics that range from medicine to ecology, from anthropology to plant and animal breeding. This book is our attempt to summarize these principles.

As noted earlier, evolution occurs as a result of successive change in the kinds and frequencies of genes that occur in populations, and it results in progressive increase in the adaptation of organisms to their environment. No organism on earth could suddenly alter the codon–amino acid correspondences in its genetic code, even though the function of a particular protein might be improved by the change, because a change in the genetic code would automatically alter every protein produced by the organism, and most of these alterations would be detrimental. In normal sexual organisms, evolution is limited by the need to work within the limits set by Mendelian segregation and recombination. Limits to evolution are also set by the patterns of control of gene activity during development; developmental pathways have evolved to function in a coordinated manner and involve so many components that they cannot undergo radical and instantaneous change. For example, in the course of human evolution there has been a dramatic increase in the size of the cranium and brain, with one result that human females sometimes have trouble giving birth because of the relatively large size of

the newborn's head. This reproductive problem cannot be circumvented completely because of the absolute necessity for intrauterine development. In short, the purpose of population genetics is to determine how evolution occurs within the framework of biological constraints.

Mathematical Models in Population Genetics

A primary tool in population genetics, as well as in such sciences as physics and chemistry, is a **model**: an intentional simplification of a complex situation, designed to eliminate extraneous detail in order to focus attention on the essentials of the situation. In population genetics, we must contend with such factors as population size, patterns of mating, geographic distribution of individuals, mutation, migration, and natural selection. Although we wish ultimately to understand the combined effects of all these factors and more, the factors are so numerous and interact in such complex ways that they cannot usually be grasped all at once. Simpler situations are therefore devised in which a few of the most important factors are identified and others are neglected. A typical laboratory experiment is a kind of model because, in an experiment, one strives to control or eliminate all sources of variation other than the variable or variables under scrutiny.

Perhaps the most important kind of model in population genetics is the **mathematical model**, which is a set of hypotheses that specifies the mathematical relationships between measured or measurable quantities (**parameters**) in a system or process. Mathematical models can be extremely useful: they express concisely the hypothesized quantitative relationships between parameters; they reveal which parameters are the most important ones in a system and thereby suggest critical experiments or observations; they serve as guides to the collection, organization, and interpretation of observed data; and they make quantitative predictions about the behavior of a system that can, within limits, be confirmed or shown to be false. The validity of a model must, of course, be tested by determining whether the hypotheses that it is based on and the predictions that grow out of it are consistent with observations.

Mathematical models are always simpler than the real world situation they are designed to elucidate. Many features of the actual system are intentionally left out of the model, because to include *every* aspect of the system would make the model too complex and unwieldy. Construction of a model always involves a compromise between realism and complexity; a completely realistic model is likely to be too complex to handle mathematically, and a model that is mathematically simple may be so unrealistic as to be useless. Ideally, a model should include all essential features of the system and exclude all

nonessential ones. How good or useful a model is often depends on how closely this ideal is approximated. In short, a model is a sort of metaphor or analogy. Like all analogies, it is valid only within certain limits and when pushed beyond these limits becomes misleading or even absurd.

The Empirical Problem

The field of population genetics has set for itself the tasks of determining how much genetic variation exists in natural populations and of explaining this variation in terms of its origin, maintenance, and evolutionary importance. Genetic variation, in the form of multiple alleles for many individual loci, exists in most natural populations; this conclusion can be documented by any number of methods, the most important of which are discussed in this chapter. On the other hand, all methods are limited to the study of a certain number of genes, usually a small number in comparison to the total number of genes in the organism. Extrapolation of the results for a small number of genes to the entire genome is questionable because of uncertainty as to whether the genes studied are truly representative of the genome. Consequently, we still have not been able to quantify how much genetic variation is "really" present in natural populations. By a variety of methods, however, we do know that the amount of genetic variation within a species, and even the nature of that variation (whether it involves DNA base substitutions, insertions, deletions, or rearrangements), varies enormously from one genetic locus to another.

Whatever total amount of genetic variation exists, sufficient variation has already been found for genes amenable to study that, for most populations, no two individuals (barring identical twins) could be expected to have the same genotype at all loci. Thus, it becomes important to describe how genetic variation in natural populations is organized into genotypes — to determine, for example, whether alleles at different loci are associated at random. As will be seen later in this chapter, the combination of particular genes into genotypes depends on the types and frequencies of matings that can occur, and on the amount of recombination between the genes.

At another level, a goal of population genetics is to understand the basis of evolutionary adaptation in populations. Often the biologist is presented with observations of features of organisms that appear to help the organism cope with its environment. Fur on a polar bear, blubber on a seal, massive incisors on predatory cats, cryptic coloration in insects, sonar in bats — the list is nearly endless and provides even the most reductionist scientist with a source of wonder and joy. Traits that give us this gut feeling as being "adaptations" to the environment can also be physiological and biochemical. The hemoglobin of many organisms has properties of oxygen affinity that

are specifically suited to the organism (Perutz 1983). Pathways of biosynthesis of plant secondary compounds to ward off herbivorous insects, and the insects' detoxification mechanisms, provide other examples. The only way we can hope to identify the genetic basis for such complex phenotypic traits is if they show variation within the species. Furthermore, in order to understand the evolution of phenotypic adaptations in populations, it is necessary to learn about the transmission of traits from parents to offspring. When traits are determined by many genes, usually in concert with environmental factors, they are referred to as **quantitative traits**, and the branch of genetics that yields predictions about similarities in quantitative traits between related individuals is **quantitative genetics**.

Around the turn of the century there was a sharp divide between two schools of thought in genetics. One school, the Mendelians, held that the sort of variation occurring in discrete classes (e.g., yellow vs. green seed color) that Mendel and others had studied was the fundamental basis of all genetically transmitted variation. They had clear, decisive experiments to support their ideas about genetic transmission. The other school, known as the biometricians, was enamored with continuous variation in quantitative traits that occurred in tiny gradations among individuals. They argued that the stuff of evolution was the transmission of this continuous sort of variation. The biometricians thought that the traits that Mendelians were studying were bizarre aberrations or "sports," and had little to do with evolution of characters such as thick fur in polar bears. These two schools were, to an extent, reconciled by the publication in 1918 of a paper by R. A. Fisher entitled "The correlation between relatives on the supposition of Mendelian inheritance." As the title implies, this was a theoretical presentation of how Mendelian transmission of many genes can result in the statistical covariance of relatives that so preoccupied the biometricians. Thus the Mendelians and biometricians were at different points on a continuum in their perceptions of the magnitude of single-gene effects.

In some ways, the rift between the Mendelians and the biometricians continues today. With the embracing of molecular techniques by many population geneticists, we finally have access to genetic variation at its most fundamental level. To some, this abundance of information has led to an embarrassment of riches: there is so much information contained in sequences that many alternative hypotheses about causes of sequence variation can be tested, sometimes with startling finality. However, sequence data removes us even further from the phenotype, and many population geneticists are concerned with adaptation at the phenotypic level. Sequence data tells us very little about a phenotype because most phenotypes are determined by many genes that are expressed through complex developmental pathways. Despite Fisher's paper, the two focuses — one on individual

genes, the other on complex phenotypes — appear to be drifting apart again. Recently, theoretical advances concerning the joint change of multiple quantitative characters have paved the way for a rebirth of interest in the evolution of quantitative characters. Although it is unlikely that any single paper will reconcile molecular and quantitative genetics any more than Fisher's paper united Mendelians and the biometricians, steady progress continues on the problems of regulation of gene expression, tissue specificity of gene splicing, homeotic mutants and other means of learning about the genetic basis of development. Molecular and quantitative population geneticists probe different problems and address different bodies of theory. However, their mutual interest in the maintenance of genetic variation and its importance to organic evolution should stimulate further productive interaction among geneticists interested in questions about populations.

DIVERSITY OF POPULATIONS

Phenotypic Variation

Most of the variation that we can readily observe in natural populations does not have a simple genetic basis. Variation in size, shape and color is nearly universal among individuals in populations, and typically no single gene accounts for the variation in traits like these. The human face is an example of ubiquitous variation that we easily accept as universal, and the critical importance of this variation and the ability it gives us to recognize one another is obvious. Morphological variation, such as occurs in facial features, is very difficult to quantify. Enormously complex statistical procedures taken from mechanical engineering have been used to formalize the description of shape and variation in shape, but these techniques are not successful in yielding a simple description. Lacking a simple description of a phenotype makes it next to impossible to find a simple genetic basis for its variation. Certain human genetic disorders result in characteristic changes in facial features (for instance, epicanthic eye folds in many individuals with the Down syndrome), but this does not mean that the great diversity in human faces has a simple genetic basis. Morphological traits are also generally affected by environmental conditions: most organisms get bigger and rounder when they eat more food. The degree to which environmental variables change morphology may vary from one genotype to another, further confounding the genetic basis for the trait.

The difficulty in finding a simple genetic basis for morphological features does not keep the evolutionary biologist from considering the importance of these traits. Many of the classic examples of adaptations, like the giraffe's neck or the camel's hump, are readily observed morphological features. The

complex genetic basis for the development of phenotype forces the population geneticist who is interested in these features to use the tools of quantitative genetics (Chapter 8), which amounts to more of a statistical description of the genetic transmission of the traits rather than identification of forces that affect the frequencies of specific genes. When we can identify single genes that have an important effect on the expression of continuously variable traits, then this type of study will help to bring together the classical, molecular and quantitative approaches in population genetics.

In addition to morphological variation, there is also continuous variation in physiological traits such as blood pressure, respiratory rate, and maximum running speed. Often physiological traits can be quantified fairly precisely, but this does not always help in identifying the genetic basis of variation. Most physiological traits are determined by many genes, but genetic lesions in single genes — even single nucleotide substitutions — can have profound effects on physiology. For example, individuals who are homozygous for the β^S globin allele have sickle cell disease, which is characterized by a complex syndrome of afflictions that relate to the inability of the blood to carry normal amounts of oxygen.

Many biochemical phenotypes are also determined by many genes and so are polygenic. Population geneticists are interested in explaining the polymorphism seen in protein sequences, and part of the explanation relates to the activity of enzymes. Individuals vary in the activity of many enzymes, including those involved in basic cell functions such as intermediary metabolism. Variation does occur in the structural genes that encode the information for the amino acid sequence of these enzymes, and this often explains differences in enzyme activity that are very large. But many other factors also affect the level of expression of the structural gene, and hence the activity of its product. For example, when genetic strains of *Drosophila* are constructed that lack variation in the structural gene, significant genetic variation among them can still be found in the activities of most of the enzymes tested (Laurie-Ahlberg et al. 1982). This type of variation is polygenic, and the appropriate tool to study it is quantitative genetics.

Simple Genetic Traits

Visible genetic variation in natural populations

The conclusion that genetic variation is widespread in natural populations is supported by many kinds of studies, particularly those demonstrating the generally deleterious effects of close inbreeding (discussed in Chapter 5), and by the success of artificial selection in altering almost any morphological or physiological characteristic (discussed in Chapter 8). Special procedures are usually required to detect genetic variation because most of it is **hidden**

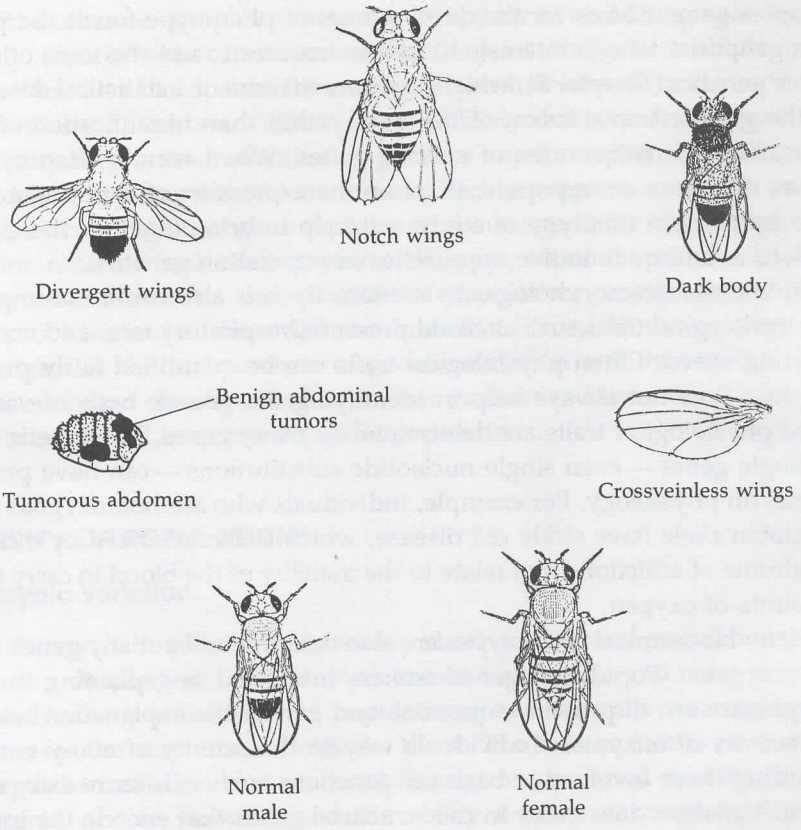


Figure 1. Examples of some phenotypic abnormalities found in natural *Drosophila melanogaster* populations; sketches of normal flies are presented for comparison. A relatively small proportion of such abnormal flies owe their abnormality to single mutant genes.

variation not apparent at the phenotypic level. Even abnormal phenotypes are difficult and sometimes impossible to trace to single genes. In humans about 5% of liveborn babies have some serious physical or mental disability, yet many of these conditions have no known genetic basis (for instance, lack of oxygen at the time of birth is an important environmental cause of mental retardation), or their mode of inheritance is polygenic and therefore complex. Similarly, about 5% of the individuals in natural populations of *Drosophila melanogaster* have phenotypic abnormalities (reviews in Spencer 1947; Lewontin 1974a). Many of the abnormalities resemble those found in special laboratory strains carrying particular mutant genes (for examples, see Figure 1), yet breeding tests indicate that no more than one-third of phenotypically

abnormal flies owe their abnormality to individual mutant genes. The rest owe their phenotypic abnormality to more complex genetic causes or to environmental accidents. Because most genetic variation is hidden, it becomes important to determine how genetic variation is organized into genotypes. This is the subject of the next section.

Protein electrophoresis

The most useful procedure yet devised for revealing genetic variation is **electrophoresis**, a technique that came into widespread use in the late 1960s (Lewontin and Hubby 1966; Harris 1966, 1969; Hubby and Lewontin 1966). One laboratory setup for electrophoresis is illustrated schematically in Figure 2. To carry out a typical electrophoresis experiment, a small sample of tissue is ground up or a small amount of blood is drawn from each of a number of individuals in the population to be surveyed. Each sample is placed in a small slot near the edge of a rectangular, thin slab of a jellylike material, usually starch or polyacrylamide. Electrodes are attached to sponge wicks soaked in buffer solution at each end of the gel, and a current is applied across the gel for several hours. Protein molecules in the sample move through the gel in response to the electric field. After electrophoresis, the gel can be stained for a particular enzyme by soaking it in a solution containing a substrate for the enzyme along with a dye that precipitates where the enzyme-catalyzed reaction occurs. A dark band thus appears in the gel, revealing the position of the enzyme. (For detailed procedures, see Harris and Hopkinson 1976.)

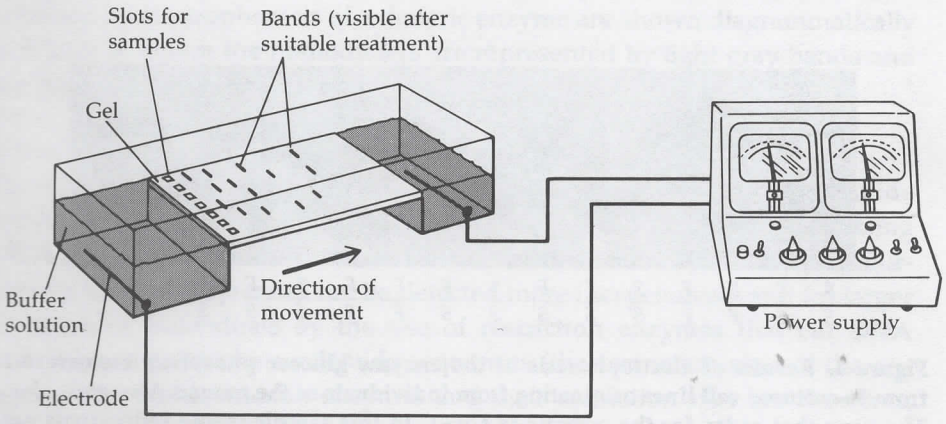


Figure 2. One type of laboratory apparatus for electrophoresis. The procedure is widely used to separate protein or DNA molecules. Starch gels (for separating proteins) and agarose gels (for separating DNA fragments) are generally run in this horizontal arrangement, while polyacrylamide gels, which are used both for separating proteins and in DNA sequencing, are typically run vertically.

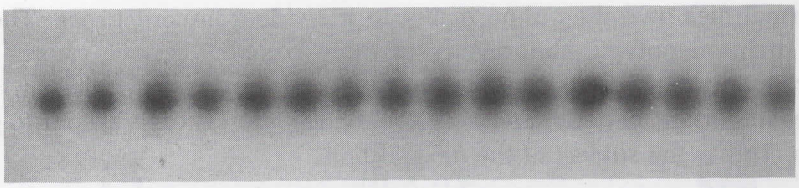


Figure 3. Results of electrophoresis of the enzyme phosphoglucomutase-1 from 16 cultured cell lines originating from individuals of the mouse, *Mus musculus*. The gene that codes for the enzyme is *Pgm-1*, and the position of the enzyme in the gel is indicated by the dark bands. In this sample, all individuals have an enzyme with the same electrophoretic mobility. (Courtesy of S. E. Lewis and F. M. Johnson.)

If an enzyme in an individual has an amino acid substitution that leads to a difference in the enzyme's overall ionic charge, then this enzyme will have an altered **electrophoretic mobility** — it will move through the gel at a different rate. This change in mobility occurs because enzymes of the same size and shape move at a rate determined largely by the ratio of the number of positively charged amino acids (primarily lysine, arginine, and histidine) to the number of negatively charged ones (principally aspartic acid and glutamic acid). Electrophoresis can therefore be used to detect mutations that result in differences in electrophoretic mobility of the corresponding enzymes.

One possible result of an electrophoresis experiment is shown in the gel photograph in Figure 3, in which all individuals in the sample have an enzyme with the same electrophoretic mobility. Another kind of result is shown in Figure 4. Here some individuals are homozygous for an allele, *F*,

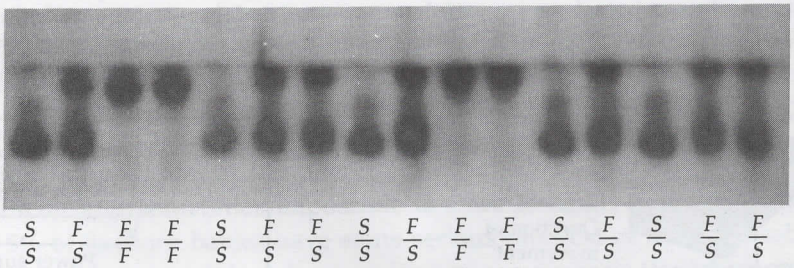


Figure 4. Results of electrophoresis of the enzyme glucose phosphate isomerase-1 from 16 cultured cell lines originating from individuals of the mouse, *Mus musculus*. The gene that codes for the enzyme is *Gpi-1*. In this sample, some individuals are homozygous for an allele (*S*) corresponding to a slow-migrating enzyme, some are homozygous for an allele (*F*) corresponding to a fast-migrating enzyme, and the rest are heterozygous (*F*/*S*). The inferred genotypes of the cell lines are indicated beneath the enzyme bands. This enzyme is a monomer, so the heterozygotes exhibit two enzyme bands of differing mobility. (Courtesy of S. E. Lewis and F. M. Johnson.)

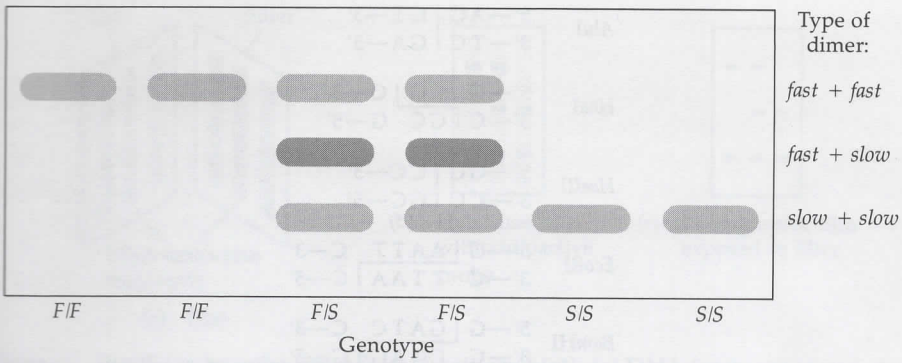


Figure 5. Typical gel pattern observed for dimeric enzymes.

associated with a rapidly migrating enzyme (*fast*); some are homozygous for an allele, *S*, associated with an enzyme that migrates more slowly (*slow*); and some are heterozygous, *F/S*, for the alleles. Two enzyme bands appear in heterozygotes because the enzyme is **monomeric**—it consists of a single polypeptide chain—and heterozygous individuals therefore produce two kinds of polypeptide chains, one corresponding to each allele. When the enzyme in question is **multimeric**, say a dimer (consisting of two polypeptide chains), then the heterozygote will produce three types of dimers: rapidly migrating (*fast* + *fast*) and slowly migrating (*slow* + *slow*) **homodimers**, like those produced in the corresponding homozygotes, and a *fast* + *slow* **heterodimer**, which typically has an intermediate electrophoretic mobility. Results obtained by electrophoresis of a dimeric enzyme are shown diagrammatically in Figure 5, where the homodimers are represented by light gray bands and the heterodimer by a dark gray band.

Restriction site variation

The ultimate resolution of genetic variation occurs at the level of nucleotide sequences in the DNA molecule, and the methods and utility in collecting DNA sequence data are discussed in the next section. However, polymorphisms in DNA sequences can be detected more inexpensively and for larger numbers of individuals by the use of **restriction enzymes** that cut DNA molecules at particular nucleotide sequences (the restriction sites of the enzymes). Several restriction enzymes and their restriction sites are shown in Figure 6. For example, the enzyme *Alu*I cuts DNA at sites containing the four-nucleotide sequence AGCT, and *Eco*RI cuts at the six-nucleotide sequence GAATTC. Most restriction enzymes used in population studies have either four- or six-nucleotide restriction sites. As we will see in Chapter 7, "four cutters" cleave the DNA much more frequently than "six cutters,"

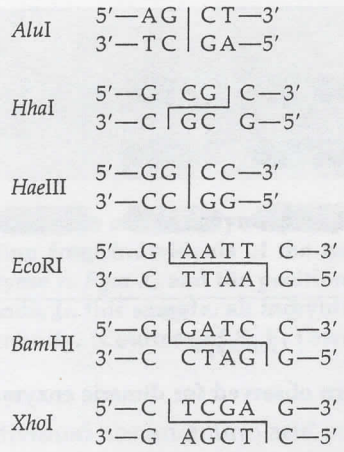


Figure 6. Restriction enzymes cleave DNA molecules at sites of specific, short nucleotide sequences. More than 500 different restriction enzymes are commercially available. They are essential tools in DNA analysis and gene cloning. The short vertical lines show the positions of cleavage in the DNA backbones.

yielding more fragments that are smaller in size. This gives four cutters an advantage in revealing more DNA sequence variation, provided the cleavage products can be interpreted.

DNA fragments can be separated by electrophoresis, in a manner similar to that used on metabolic enzymes. Unlike metabolic enzymes, which can be identified by their unique enzymatic activity, the identification of DNA fragments requires special procedures. The most widely used of these procedures is the Southern blot (Southern 1975), illustrated in Figure 7. The key element in the Southern blot is the use of a **probe DNA molecule** containing the nucleotide sequences of interest, which is obtained from a gene that has been cloned into bacterial cells. In the Southern procedure, DNA restriction fragments that have been separated by electrophoresis are made single stranded by soaking the gel in an alkaline solution. The single-stranded DNA fragments are then blotted onto a nitrocellulose or nylon filter, where either heat or ultraviolet light is used to attach the DNA to the filter (Figure 7). The filter is then prehybridized with herring or salmon sperm DNA to prevent nonspecific binding of the probe DNA to the filter. The filter is next bathed in a solution containing single-stranded probe DNA that has been rendered radioactive (one process for doing this is called **nick translation**). The probe DNA forms double-stranded molecules with its complementary counterparts on the filter. Careful washing removes excess probe DNA that has remained unpaired. The filter is then sandwiched with photographic film, where radioactive disintegrations from the bound probe on the filter result in visible bands on the film.

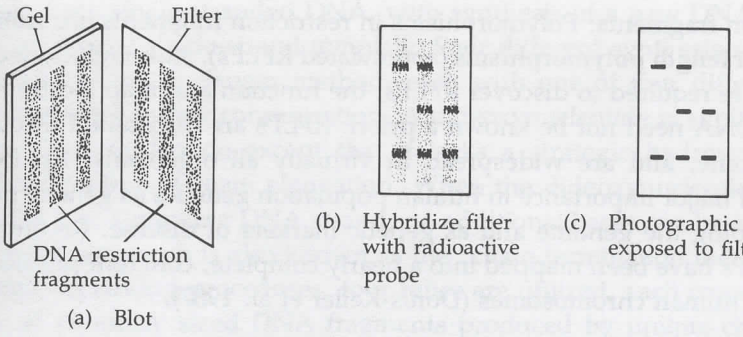


Figure 7. Southern transfer procedure (Southern blot). (a) DNA fragments separated by electrophoresis are transferred and chemically attached to a filter. (b) The filter is mixed with radioactive probe DNA, which sticks to homologous DNA molecules in the filter. The drawing shows where the probe binds, although it is not visible at this stage. (c) After washing, the filter is exposed to photographic film, which develops dark bands caused by radioactive emissions from the probe.

Polymorphisms resulting in the presence or absence of restriction sites can be identified because they change the length of characteristic restriction fragments (Figure 8). The upper part of each panel shows the location of restriction sites in the DNA molecules in a diploid individual; the *a*-type molecule contains one additional restriction site that is not present in the *A*-type molecule. The lower part of the figure demonstrates that, with suitable probe DNA, all three genotypes can be distinguished by their pattern of

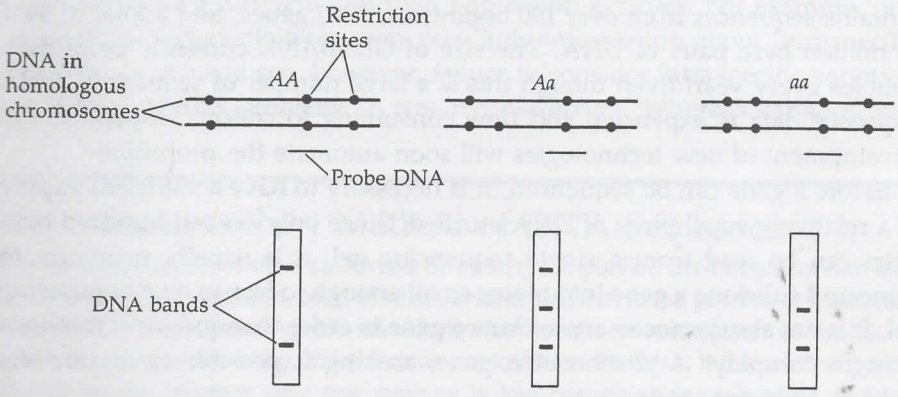


Figure 8. Restriction fragment length polymorphisms (RFLPs) result from the presence or absence of particular restriction sites in DNA. The DNA molecule designated *A* contains three restriction sites; the one designated *a* contains four. Genotypes *AA*, *Aa* and *aa* each give a different pattern of bands in Southern transfer experiments using the indicated probe DNA.

restriction fragments. Polymorphisms in restriction fragments are **restriction fragment length polymorphisms** (abbreviated RFLPs). Although cloned DNA probes are required to discover RFLPs, the function and map location of the cloned DNA need not be known a priori. RFLPs are distributed throughout the genome, and are widespread in virtually all organisms. RFLPs have assumed major importance in human population genetics as genetic markers for mapping the genome and as genetic markers of disease. Recently over 400 RFLPs have been mapped into a nearly complete, coherent physical map of all 23 human chromosomes (Donis-Keller et al. 1987).

DNA sequence variation

Nucleotide sequences of genes yield the ultimate level of resolution of genetic variation. If the protein bands from two individuals migrate the same way in protein electrophoresis, or the restriction fragment patterns of the genes look the same, we still cannot be sure that the alleles are identical. However, if the DNA sequences are identical, then so are the genes. It is highly unlikely that two genes identical in DNA sequence derive from different ancestral genes. This principle of DNA sequence identity is particularly important for many current tests about evolutionary forces that make use of DNA sequence data.

DNA sequences contain an impressive amount of information, and an important resource for molecular biologists is to be able to search through data banks of known DNA sequences for similarities to newly sequenced genes. Several such data banks have been organized, including one by the European Molecular Biology Laboratory (EMBL) and one by the U.S. National Institutes of Health (NIH) called GENBANK. At this writing, GENBANK contains sequences from over 100 organisms, 9000 genes, and a total of over 20 million base pairs of DNA. The size of GENBANK currently more than doubles every year! Even though this is a large number of sequences, DNA sequence data is expensive and time consuming to collect. Hopefully, the development of new technologies will soon automate the procedure.

Before a gene can be sequenced, it is necessary to have a sufficient supply of a relatively small piece of DNA purified. Since only several hundred base pairs can be read from a single sequencing gel, it is usually necessary to clone and subclone a gene into pieces small enough to be run on a sequencing gel. It is not always necessary to clone a gene in order to sequence it; methods exist to "amplify" a portion of a gene, making it possible to accumulate enough DNA for sequencing.

There are two primary methods for sequencing DNA. The first, known as Maxam-Gilbert sequencing, uses DNA samples that are radioactively labeled at the 3′ end and produces breaks at the 5′ end by reactions that are specific to different nucleotides. The other method, known as Sanger dideoxy se-

quencing, uses single-stranded DNA, with synthesis of a new DNA strand carried out on the single-strand template. Four different synthesis reactions are carried out in the Sanger method, each with one of four different dideoxynucleotides in low concentration. A dideoxynucleotide is similar to an ordinary deoxynucleotide except that it lacks a strategic hydroxyl group needed for continued chain elongation. When the dideoxynucleotide is incorporated into a growing DNA strand, no additional bases can be added to it; hence this approach is also known as the "chain termination method."

In both sequencing procedures, four tubes are utilized, each containing a mixture of variously sized DNA fragments produced by unique chain-terminating events. The fragments reproduced in each tube are separated according to size by electrophoresis in a special acrylamide gel, which enables separation of short DNA fragments that differ by as little as one nucleotide in length. After the gel is run, it is dried and placed in contact with X-ray film for autoradiography. Small fragments of DNA that were radioactively labeled before the electrophoretic separation appear as dark bands on the autoradiograph. Since there are different arrays of fragment lengths in the four lanes that are run for each segment of DNA, the band pattern gives the DNA sequence directly. (See Box A for an example of Maxam-Gilbert sequencing. Box B in Chapter 7 provides an example of the Sanger method.)

Most DNA sequencing is done for the purpose of identifying the nature of gene regulation: regulatory elements are identified and hypotheses about this regulation can be directly tested by modifying the DNA sequence. Some investigators are interested in the evolutionary relationships among organisms, and DNA sequence data can be used for phylogenetic reconstruction. The inference of phylogenies requires evolutionary reasoning, but phylogenetics has quite a different aim than population genetics. For example, one can draw phylogenetic trees with very little information about intraspecific variation. As we will see, however, failure to consider intraspecific variation can lead to serious problems in tree reconstruction. Whereas variation is a

The Maxam-Gilbert Method of DNA Sequencing

A

Maxam-Gilbert sequencing is performed by treating the purified DNA fragment with four different chemical reactions, each of which breaks the DNA at the position of a different base. For example, treating with dimethylsulfate results in methylation of the N_7 of guanines, and upon heating, the guanines are lost from the DNA. If this is done gently so that on the average only one guanine is lost per molecule, a number of DNA fragments will be generated, each ending where there was formerly a guanine. For example, the sequence ATGTTAGCCGCTA*, where * represents the radioactively labeled 3′ end, would yield fragments CTA*, CCGCTA* and TTAGCCGCTA*. By treating with dimethylsulfate in combination with acid, the breakage occurs at adenines as well

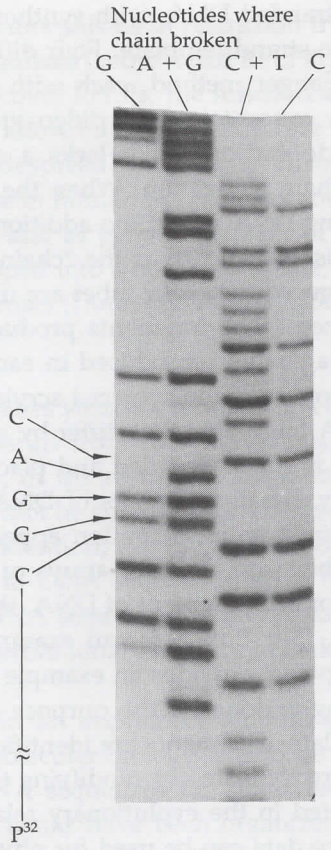


Figure 9. In the Maxam-Gilbert sequencing method, a single DNA strand is radioactively labeled at one end and gently treated with four types of chemical reagents that break the chain at particular combinations of nucleotides. When the resulting fragments are separated in a gel, the nucleotide present at any position in the original DNA can be determined from which chemical treatment caused breakage at that position. Some of the reagents are not specific for individual nucleotides; in this example, the reagents break at G (lane 1), A and G (lane 2), C and T (lane 3), or C (lane 4). Each nucleotide in the chain gives a unique pattern. For example, a G gives bands in lanes 1 and 2, but an A gives a band in lane 1 only. The diagram at the right shows how the nucleotide sequence CGGAC is determined from the gel. Nucleotides 11–35 in this molecule are CGGACGTCGTCTTTACCAACTTGAG. (Courtesy of John Majors.)

as guanines; by treating with hydrazine and salt the breakage occurs at cytosines; and by treating with hydrazine alone the breaks occur at both cytosine and thymine. After these four reactions are carried out the resulting fragments are separated on a thin polyacrylamide gel. See if you can read the sequence of the first 10 nucleotides from the gel depicted in Figure 9.

ANSWER Starting at the *bottom* of the gel, the sequence is TTTACGGCGC.

bother to the phylogeneticist, it is the focus of interest to the population geneticist. The ideal DNA sequence data for a population geneticist is a collection of sequences of several alleles from many individuals of the same population of the same species. Generating such data obviously entails a great deal of very tedious work, but the payoff can be quite rewarding. For example, the degree of polymorphism for sites that do not result in amino acid changes (silent sites) can be compared to the degree of amino acid polymorphism, or variation in structural parts of genes can be compared to variation in flanking regions (some of which might be regulatory) and in introns. An example of the power of this technique will be briefly presented in the section on molecular sequence variation (p. 25).

ALLELE FREQUENCIES AND POLYMORPHISM

In order to compare different genes and different populations, it is necessary to have some convenient quantitative measure of genetic variation. Genetic variation can be quantified using the concept of allele frequency. The **allele frequency** of a prescribed allele among a group of individuals is simply the proportion of all alleles of the gene that are of the prescribed type. The frequency of any prescribed allele in a sample is therefore equal to twice the number of homozygotes for the allele (because each homozygote carries two copies of the allele) plus the number of heterozygotes for that allele (because each heterozygote carries one copy) divided by two times the number of individuals in the sample (because each individual carries two alleles of the gene). Among the small sample of 16 individuals in the gel shown in Figure 4, for example, there are 4 *F/F* homozygotes, 7 *F/S* heterozygotes, and 5 *S/S* homozygotes; the allele frequency of *F* is thus $(2 \times 4 + 7)/(2 \times 16) = 0.469$, whereas the allele frequency of *S* is $(2 \times 5 + 7)/(2 \times 16) = 0.531$. (Detailed calculations are shown in Table 1.) Note that the sum of all allele

Table 1. Calculation of allele frequencies.

	GENOTYPE			
	F/F	*F/S*	*S/S*	**Total**
Number of individuals	4	7	5	16
Number of *F* alleles	8	7	0	15
Number of *S* alleles	0	7	10	17
Number of *F* + *S* alleles	8	14	10	32
Allele frequency of *F* = 15/32 = 0.469				
Allele frequency of *S* = 17/32 = 0.531				

frequencies, in this case those for *F* and *S*, must be 1. The allele frequency in a sample of individuals from a population is only an estimate of the true allele frequency in the whole population, but the estimate will usually be close to the true frequency if the sample is sufficiently large and representative; for this reason, allele frequency estimates should be based on samples of 100 or more individuals. One further comment about allele frequency: an allele frequency is often imprecisely called a **gene frequency** if the allele in question is clear from context. (See Box B for calculations of allele frequency involving actual data.)

B Calculating Allele Frequencies

Mukai, Watanabe and Yamaguchi (1974) studied 1158 second chromosomes from a population of *Drosophila melanogaster* in Raleigh, North Carolina. Two alleles affecting electrophoretic mobility were found at the α-glycerophosphate dehydrogenase-1 locus, denoted α-*GPDH*-1^F (for the allele determining the fast migrating allozyme) and α-*GPDH*-1^S (for the slow form). Two electrophoretic alleles were also found at the α-amylase locus, denoted Amy^F and Amy^S. The chromosomes were also classified according to whether they carried the polymorphic paracentric inversion called the *NS* (Nova Scotia) inversion on the right arm of the chromosome. Results were as follows:

α-*GPDH*-1^F	Amy^F	non-*NS*	726
α-*GPDH*-1^F	Amy^F	*NS*	90
α-*GPDH*-1^F	Amy^S	non-*NS*	111
α-*GPDH*-1^F	Amy^S	*NS*	1
α-*GPDH*-1^S	Amy^F	non-*NS*	172
α-*GPDH*-1^S	Amy^F	*NS*	32
α-*GPDH*-1^S	Amy^S	non-*NS*	26
α-*GPDH*-1^S	Amy^S	*NS*	0

Calculate the frequencies of the α-*GPDH*-1^F and α-*GPDH*-1^S alleles. Calculate the frequencies of the Amy^F and Amy^S alleles. Calculate the frequencies of the non-*NS* and *NS* inversion types.

ANSWER The frequency of the α-*GPDH*-1^F allele is (726 + 90 + 111 + 1)/1158 = 0.8014, and the other allele at this locus has a frequency of 0.1986. The frequency of the Amy^F allele is (726 + 90 + 172 + 32)/1158 = 0.8808, and the Amy^S allele has a frequency of 0.1192. The *NS*-bearing chromosome frequency is (90 + 1 + 32 + 0)/1158 = 0.1062, and the non-*NS*-bearing chromosome frequency is 0.8938.

The idea of allele frequency supports the notion of polymorphism. A **polymorphic gene** is a gene at which the most common allele has a frequency of less than 0.95. Conversely, a **monomorphic gene** is one that is not poly-

morphic. The cutoff at 0.95 in the definition of polymorphism is arbitrary, but it serves to focus attention on those genes with common allelic variation. If you look long enough and hard enough in any large population, you are bound to find **rare alleles** — alleles with frequencies of less than 0.005 — of virtually every gene. In humans, between one and two individuals per thousand are heterozygous for rare alleles at any locus, as judged by an electrophoretic enzyme survey of 43 genes in 250,000 Europeans (Harris and Hopkinson 1972). Many rare alleles are deleterious and are presumably maintained in populations by recurrent mutation, as will be discussed in Chapter 3. The definition of polymorphism is an attempt to focus on genes having alleles with frequencies too high to be explained solely by recurrent mutation. With the definition of polymorphism given above, at least 2% of the population will be heterozygous for the most common allele, for reasons that will soon become clear.

ALLOZYME POLYMORPHISMS

Enzymes differing in electrophoretic mobility as a result of allelic differences in a single gene are called **allozymes**. Thus, allozyme variation in a population is an indication of genetic variation, and it turns out that such genetic variation is very common. Extensive allozyme variation has been found in virtually all natural populations studied by electrophoresis, including the bacterium *Escherichia coli* (Selander and Levin 1980; Selander et al. 1987), plants (for representative examples, see Clegg and Allard 1972; Hamrick and Allard 1972; Levin 1978), *Drosophila* (see the classic paper by Lewontin and Hubby 1966, and those by Richmond 1972; Prakash 1977; Marinkovic et al. 1978), the mouse (e.g., Selander et al. 1969), and humans (the classic study is by Harris 1966, but see also Ruddle et al. 1969; Harris and Hopkinson 1972).

Polymorphism for alleles that determine allozymes is extremely widespread. This is shown in Figure 10, which summarizes the results of electrophoretic surveys of up to 71 genes in populations of 243 species (Nevo 1978). Each point in the figure (except for the one pertaining to humans) gives the type of organism studied and the number of species examined. The vertical axis, labeled $\langle P \rangle$, refers to the estimated proportion of genes that are polymorphic by the 0.95 criterion; and the horizontal axis, labeled $\langle H \rangle$, refers to the estimated proportion of genes expected to be heterozygous in an average individual. The average heterozygosity H is estimated by totaling the number of heterozygous individuals for each gene, dividing this by the total number of individuals in the sample, and averaging over all genes. To illustrate calculation of P and H, we may use the data of Harris (1966), who studied

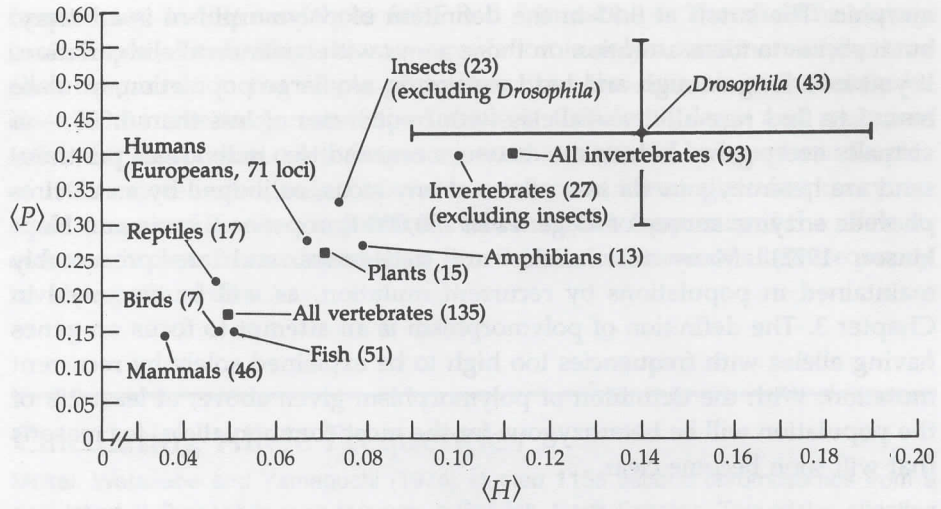


Figure 10. Estimated levels of heterozygosity ($\langle H \rangle$) and proportion of polymorphic genes ($\langle P \rangle$) derived from allozyme studies of various groups of plants and animals. The number of species studied is shown in parentheses beside each point. Squares denote averages for plants, invertebrates, and vertebrates. The bars across the *Drosophila* point show the range of H and P within which 68% of the *Drosophila* species fall. Other groups would have similarly large bars. (Data from Nevo 1978.)

ten enzyme genes in the English population. Of these ten genes, three were found to be polymorphic, so in this sample $\langle P \rangle = 3/10 = 0.33$. The proportion of heterozygous individuals for each of the seven monomorphic genes was 0, and the proportion of heterozygous individuals for the three polymorphic loci was 0.509 (for red cell acid phosphates), 0.385 (for phosphoglucomutase), and 0.095 (for adenylate kinase); thus, in this sample

$$\langle H \rangle = (7 \times 0 + 0.509 + 0.385 + 0.095)/10 = 0.099$$

The vertical and horizontal bars on the point corresponding to *Drosophila* indicate substantial variability found in P and H among species in a group. The bars indicate the limits of P and H within which 68% of the individual species are expected to fall. Thus, among *Drosophila* species, 68% are expected to have a proportion of polymorphic genes in the range 0.30 to 0.56, and an average heterozygosity in the range 0.09 to 0.19. Such bars could be attached to each point, and their lengths would be comparable to those for *Drosophila*.

It is difficult to summarize concisely the results of Figure 10 because of the immense variability in polymorphism and heterozygosity found within each group of organisms, as indicated by the length of the variability bars corresponding to *Drosophila*. On the whole, there is a positive relationship between the amount of polymorphism and the degree of heterozygosity,

which is as expected because the greater the fraction of polymorphic genes in a population, the more genes are expected to be heterozygous in an average individual. The overall mean of $\langle P \rangle$ in Figure 10 is 0.26 ± 0.15, and the mean of $\langle H \rangle$ is 0.07 ± 0.05. Vertebrates have the lowest average amount of genetic variation among the groups, plants come next, and invertebrates have the highest. *Drosophila* is the most genetically variable group of higher organisms so far studied, and mammals the least variable. Humans are fairly typical of large mammals, and an extensive electrophoretic survey of 104 genes in a sample including all major human races gave estimates of $\langle P \rangle = 0.32$ and $\langle H \rangle = 0.06$ (Harris et al. 1977). The one obvious conclusion that can be reached from Figure 10 is that allozyme polymorphisms are widespread among higher organisms. Genetic variation is even more prevalent among some prokaryotes. For example, natural isolates of the mammalian intestinal bacterium *Escherichia coli* exhibit levels of genetic polymorphism two or three times greater than occur in species of eukaryotes (Selander and Levin 1980; Whittam et al. 1983). Box C is an example of "raw data" collected in an allozyme survey of a hypothetical population.

Allozyme Data: A Simple Example

C

The gel patterns shown in Figure 11 are the results of electrophoresis of a random sample of 50 individuals from a hypothetical population. Tissue samples from each of the individuals were electrophoresed, and gel slices were stained for five enzymes corresponding to slices A through E. Breeding tests demonstrated that differences in enzyme mobility were in each case due to alleles of a single gene. The population is diploid and undergoing random mating with respect to these genes. All five enzymes are either monomers or dimers.

a. Which enzymes are monomeric, which dimeric, and which are not clear from these gels?
b. How many electrophoretically distinct alleles are there for each of the genes in gels A through E?
c. What are the allele frequencies at each locus?
d. Which of the genes are polymorphic in this sample? What is the proportion of polymorphic genes?
e. What is the average heterozygosity for each gene? What is the average heterozygosity for all five genes?

ANSWER **a.** The monomers are the enzymes on gels A, B and E; D is a dimer, and we cannot tell from C what the enzyme structure is.
b. Gels A, D, and E reveal two alleles (call them *F* and *S*, for fast and slow); gel B shows three alleles (call them *F*, *S*, and *I*, for intermediate); and gel C shows no variation, or one allele.
c. The genotypes A^FA^F, A^FA^S and A^SA^S in gel A number 32, 16 and 2 individuals, so

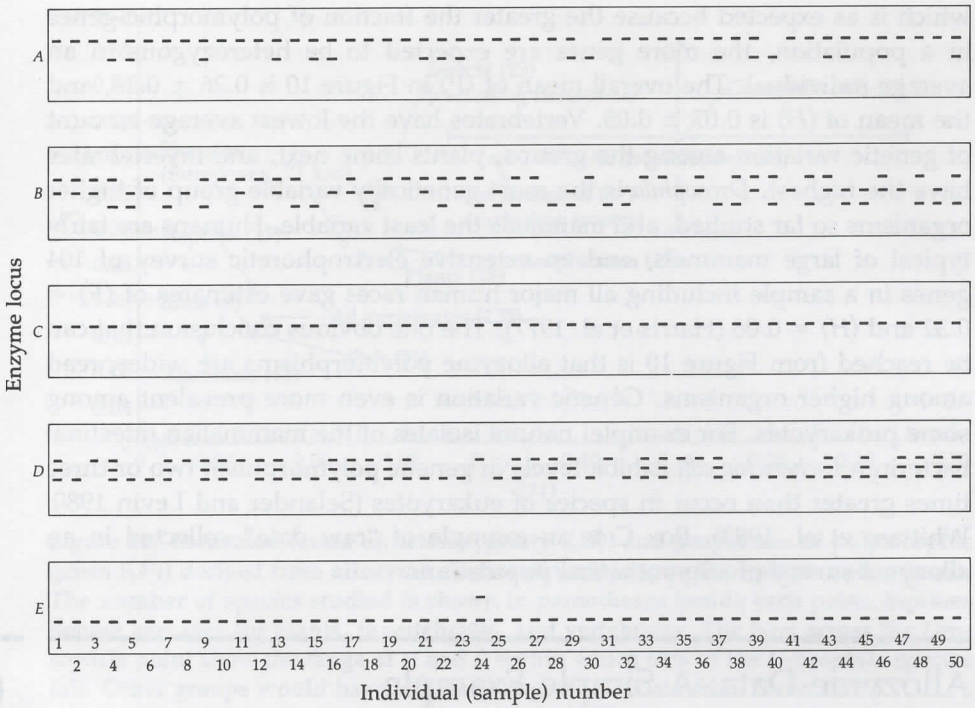


Figure 11. Results of a hypothetical population survey using protein electrophoresis. Each of the 50 lanes represents a different individual, and the 5 slabs designated A through E represent horizontal slices of the gel stained to indicate 5 different enzymes. (Refer to the problem in Box C.)

the allele frequency of *F* is (64 + 16)/100 = 0.8, and the *S* allele has frequency 0.2. In B, the genotypes and their counts are B^FB^F (7), B^FB^I (13), B^IB^I (5), B^IB^S (9), and B^SB^S (4), so the allele frequencies are, for B^F: (14 + 13 + 12)/100 = 0.39; for B^I: (13 + 10 + 9)/100 = 0.32; and for B^S: (1 − 0.39 − 0.32) = 0.29. The one allele on gel C has frequency 1.0. The genotypes and their counts for gel D are D^FD^F (8), D^FD^S (24) and D^SD^S (18), so the allele frequencies are, for D^F: (16 + 24)/100 = 0.4; and for D^S: 0.6. The genotype frequencies and their counts for the E gel are E^FE^F (0), E^FE^S (1), and E^SE^S (49), so the allele frequencies are, for E^F: (0 + 1)/100 = 0.01; and for E^S: (0.99).

d. Genes A, B and D are polymorphic, and C and E are considered monomorphic because the frequency of the most common allele is greater than or equal to 0.95. The proportion of loci that is polymorphic is therefore $P = 3/5 = 0.6$.

e. The frequency of heterozygotes (or the heterozygosity) of each gene is: H_A, 0.32; H_B, 0.68; H_C, 0; H_D, 0.48; and H_E, 0.02. The average heterozygosity is H = (0.32 + 0.68 + 0 + 0.48 + 0.02)/5 = 0.30.

Although genetic polymorphisms are widespread, they are not universal. For example, both major subspecies of the cheetah (*Acinonynx jubatus*) are

virtually monomorphic. A survey of 49 enzymes among 30 individuals from the East African subspecies (*A. jubatus raineyi*) yielded only two polymorphic genes and estimates of $\langle P \rangle = 0.04$ and $\langle H \rangle = 0.01$; among 98 individuals from the South African subspecies (*A. jubatus jubatus*), the estimates were $\langle P \rangle = 0.02$ and $\langle H \rangle = 0.0004$ (O'Brien et al. 1987). Most unusual was the finding of skin graft acceptance between unrelated cheetahs from the South African subspecies. Graft acceptance means that the cheetah population is monomorphic for the major histocompatibility locus, which is abundantly polymorphic in other mammals. The cheetah was worldwide in its range at one time but presently is found only in Africa, where there are fewer than 20,000 animals. Apparently, the cheetah underwent at least two severe constrictions in population number, resulting in the loss of most of its genetic variability (O'Brien et al. 1987).

The validity of estimates of polymorphism based solely on protein electrophoresis is open to question. The amount of polymorphism may be underestimated because routine electrophoresis fails to detect many amino acid substitutions. The resolving power of electrophoresis can be enhanced by using gels of varying ionic concentrations or pH conditions (**sequential electrophoresis**); proteins not separated at one ionic concentration or pH may be separated at another (Singh et al. 1976; Ramshaw et al. 1979). Using five pH conditions, McLellan and Inouye (1986) were able to distinguish 13 of 14 myoglobins known to be different. Some amino acid substitutions that do not alter electrophoretic mobility do render the enzyme sensitive to high temperature (Bernstein et al. 1973; Trippa et al. 1976). Incubating the sample at high temperature prior to staining may destroy enzyme activity. The diagrammatic example in Figure 12 shows gels stained for enzyme activity (from two samples) with or without heat treatment. One enzyme is seen to be temperature sensitive, indicating that it contains one or more amino acid

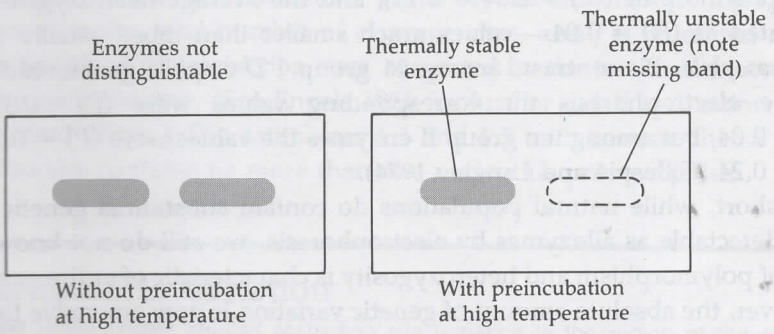


Figure 12. Expected gel patterns resulting from heat treatment of a thermally unstable (thermolabile) enzyme. The gel on the right has been incubated at high temperature before staining and the thermolabile enzyme has been rendered inactive.

substitutions not found in the other enzyme. The use of techniques of this sort has increased the number of identified alleles at the *xanthine dehydrogenase* locus in *Drosophila pseudoobscura* from 6 to 37, and increased estimates of average heterozygosity from 0.44 to 0.73 (Singh et al. 1976). Greater richness of allelic diversity of *Xdh* and *Esterase-5* has proven very useful for testing the goodness-of-fit of population genetic models, and for making inferences about population subdivision (Keith 1983; Keith et al. 1985). Using sequential electrophoresis, Bonhomme and Selander (1978) were able to identify about twice as many alleles in the house mouse as had been previously identified by conventional electrophoresis. On the other hand, while more precise electrophoretic techniques reveal additional alleles of genes known to be polymorphic, thus increasing estimates of H, genes classified as monomorphic by means of routine electrophoresis tend to remain monomorphic even with more critical techniques (Coyne et al. 1978)—so estimates of P remain much the same.

Although routine electrophoretic surveys may, for the reasons given above, miss a significant amount of genetic variation, they could also overestimate the amount of polymorphism, because the enzymes typically surveyed are those found in relatively high concentration in tissues or body fluids. Such enzymes are often called **group II enzymes** to distinguish them from more substrate-specific **group I enzymes** involved in processes such as energy transformation (Gillespie and Kojima 1968). Group II enzymes are hardly a random sample of genes because the sample excludes not only group I enzymes, but also regulatory genes and genes that code for the several dozen ribosomal proteins and transfer RNAs. Group II enzymes are commonly used in electrophoretic population surveys because they are the easiest to assay. Leigh Brown and Langley (1979) have used a powerful two-dimensional electrophoretic technique to study the 54 proteins found in highest concentration in *Drosophila*; these 54 proteins included only six that were polymorphic ($\langle P \rangle = 6/54 = 0.11$), and the average heterozygosity was estimated as $\langle H \rangle = 0.04$—values much smaller than those usually quoted for *Drosophila*. By contrast, among 11 group I *Drosophila* genes studied by routine electrophoresis the corresponding values were $\langle P \rangle = 0.27$ and $\langle H \rangle = 0.04$, but among ten group II enzymes the values were $\langle P \rangle = 0.70$ and $\langle H \rangle = 0.24$ (Gillespie and Langley 1974).

In short, while natural populations do contain substantial genetic variation, detectable as allozymes by electrophoresis, we still do not know what level of polymorphism and heterozygosity is characteristic of entire genomes. However, the absolute amount of genetic variation is now amenable to study at the DNA level, either directly by means of DNA sequencing or indirectly by examining variation in restriction sites. This is the subject of the next section.

The amount of polymorphism at the nucleotide level can be estimated indirectly from restriction site data. For example, suppose a study includes n homologous DNA molecules, each from a different individual in a population or from individuals of different species; these are cut with a restriction enzyme and found to contain some number m of cleavage sites (sites at which cleavage occurs in at least one molecule in the sample). For cleavage site i, let c_i represent the number of molecules in the sample which are cleaved. Thus, for $i = 1, 2, \ldots, m$, the particular value of c_i may range from 1 to n. The quantity $c = \Sigma c_i$ equals the total number of cuts at all cleavage sites in the sample. Estimates of the proportion of polymorphic nucleotide sites in the sample P_{nuc} and the heterozygosity per nucleotide site in the sample H_{nuc} have been determined by Engels (1981) as

$$\langle P_{\text{nuc}} \rangle = \frac{c - n(m - k)}{jc} \tag{1.1}$$

$$\langle H_{\text{nuc}} \rangle = \frac{nc - \Sigma c_i^2}{jc(n - 1)} \tag{1.2}$$

where k is the number of cleavage sites that are absent in at least one member of the sample, and j is the number of nucleotides in the restriction site (usually either 4 or 6). The subscript "nuc" indicates that polymorphism and heterozygosity are estimated at the level of individual nucleotides. The estimate of H obtained for allozymes is the *observed* heterozygosity, and is simply the fraction of the population that is heterozygous averaged over loci. H_{nuc} is not calculated as observed heterozygosity; rather, Equation 1.2 is the *expected* heterozygosity obtained from a particular model. The reason expected heterozygosity is used is that the data are assumed to be from a single DNA molecule from each individual. When several enzymes are used, m, k, and c_i refer to the total number of cleavage sites, and j is redefined as the average length of the restriction sites, weighted by the total number of cuts made by each enzyme. (See Engels 1981 for further detail and estimates of the standard errors.) The estimators 1.1 and 1.2 both assume that a given restriction site contains no more than one nucleotide polymorphism.

Restriction Site Variation

D

Cross and Birley (1986) studied restriction site variation in the region of the alcohol dehydrogenase gene of *Drosophila melanogaster* in a population descended from animals trapped at a Dutch fruit market in Gröningen. The region contained a total of 23 sites for 5 restriction enzymes, each having a 6-base restriction site, and 16 of the

sites were cut in all 10 individuals in the sample. The data below document the presence (+) or absence (−) of the 7 sites found to be polymorphic in a sample of 10 chromosomes. Use Equations 1.1 and 1.2 to estimate the proportion of polymorphic nucleotides and the heterozygosity of nucleotide sites in the sample.

Individual	*Bam*H1	*Hind*III	*Pst*I	*Xho*I	*Pst*I	*Eco*RI	*Eco*RI
1	+	−	−	+	+	−	−
2	+	−	−	−	−	+	+
3	−	−	+	−	−	+	−
4	−	+	−	+	−	+	+
5	−	+	−	+	−	+	+
6	−	+	−	+	−	+	+
7	−	+	−	+	−	+	+
8	−	+	−	+	−	+	+
9	−	+	−	+	−	+	+
10	−	−	−	+	−	+	−

ANSWER For this sample, $n = 10$, $m = 16 + 7 = 23$, $k = 7$ and $j = 6$. Taking the sites from left to right, $c_1 = 2, c_2 = 6$, $c_3 = 1$, $c_4 = 8$, $c_5 = 1$, $c_6 = 9$, $c_7 = 7$, and there are 16 additional sites for which $c_i = 10$. Thus, $c = \Sigma c_i = 34 + 160 = 194$ and $\Sigma c_i^2 = 236 + 1600 = 1836$. From Equation 1.1, $\langle P_{nuc} \rangle = [194 - 10(23 - 7)]/(6)(194) = 0.029$ and $\langle H_{nuc} \rangle = [(10)(194) - 1836]/(6)(194)(9) = 0.010$. (The actual sample of 34 chromosomes gave almost the same estimates.)

The variation detected as RFLPs has proven useful in two distinct ways. First, by constructing a restriction map, so that the locations of the restriction sites are all known relative to each other, it is possible to represent alleles as vectors of plus and minus signs indicating the presence or absence of restriction sites (known as a **haplotype**), as the example in Box D demonstrates. The study of restriction site haplotypes has an advantage over protein electrophoresis because many regions of the genome are sufficiently variable that many haplotypes can be identified. In humans, a noninbred individual is heterozygous at about one nucleotide site in every 200, and this is also about the frequency of sites that differ between two randomly chosen individuals. Thus, the method has greater resolution than enzyme electrophoresis in estimating genetic similarity among individuals; this has value in determining relationships among individuals, and in testing different models that attempt to explain levels of observed variation. RFLP analysis is much more expensive than protein electrophoresis, and as a consequence, sample sizes are generally smaller. This results in larger sampling variances for estimates of allele frequencies and disequilibria. RFLPs are generally identified by using the Southern blotting procedure. Some probes identify sequence variation at extremely variable sites known as **minisatellites**. Because of the high level

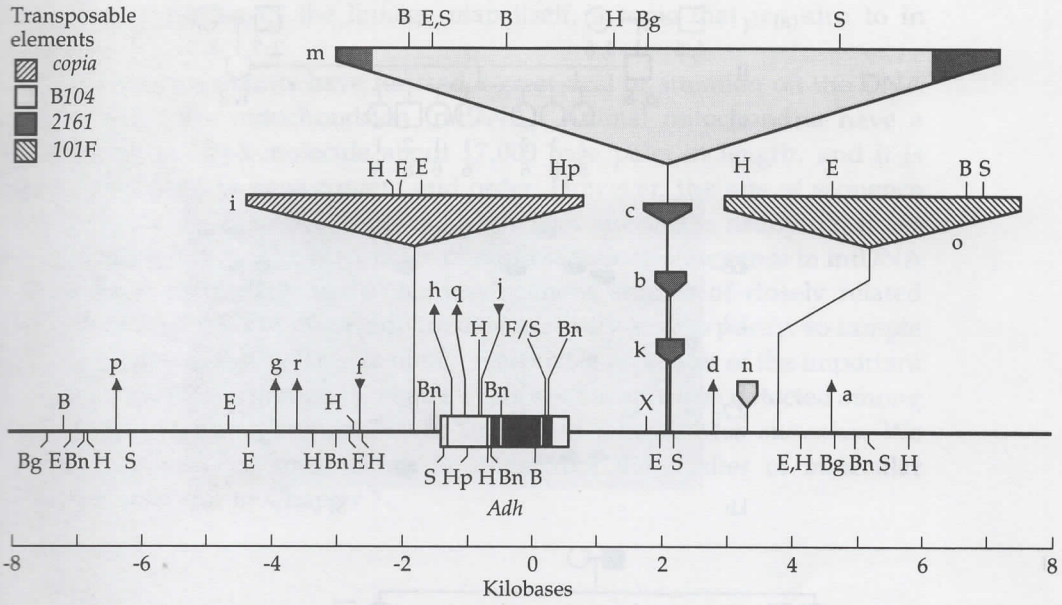


Figure 13. Restriction site map of naturally occurring DNA sequence variation in the *Adh* region of *Drosophila melanogaster*. The data for the map were obtained by the Southern transfer procedure on a series of homozygous strains of *D. melanogaster*. The boxed region in the center of the map indicates the *Adh* gene. Downward pointing triangles indicate insertions, the largest of which are transposable elements (identified in the key). Upward pointing triangles indicate small deletions. The unique insertions and deletions are represented by solid black triangles. Variant and invariant restriction sites are indicated above and below the line, respectively, as follows: B(*Bam*HI), Bg (*Bgl*II). Bn (*Ban*II), E (*Eco*RI), H (*Hin*dIII), Hp (*Hpa*I), S (*Sal*I), and X (*Xho*I). With minor exceptions, each of these sites was scored for all 49 strains. (From Aquadro et al. 1986.)

of variation of minisatellites, this method is often referred to as "DNA fingerprinting;" it is useful in paternity analysis, both in humans and in natural populations of many organisms. Finally, RFLP analysis has revealed surprises — such as extensive variation due to insertions and deletions of DNA segments, including that of transposable elements (Figure 13).

The second major use of RFLP variation has been in the construction of genetic linkage maps of the genomes of several organisms, including humans, yeast, *Escherichia coli*, and the nematode *Caenorhabditis elegans*. These maps are especially useful for locating genes that are responsible for genetic disorders, and they yield many interesting insights for the population geneticist as well. Figure 14 depicts the linkage between the gene for a genetic disorder and a nearby RFLP, and demonstrates that the presence or absence of restriction sites obeys Mendel's law of segregation. Based on RFLP map-

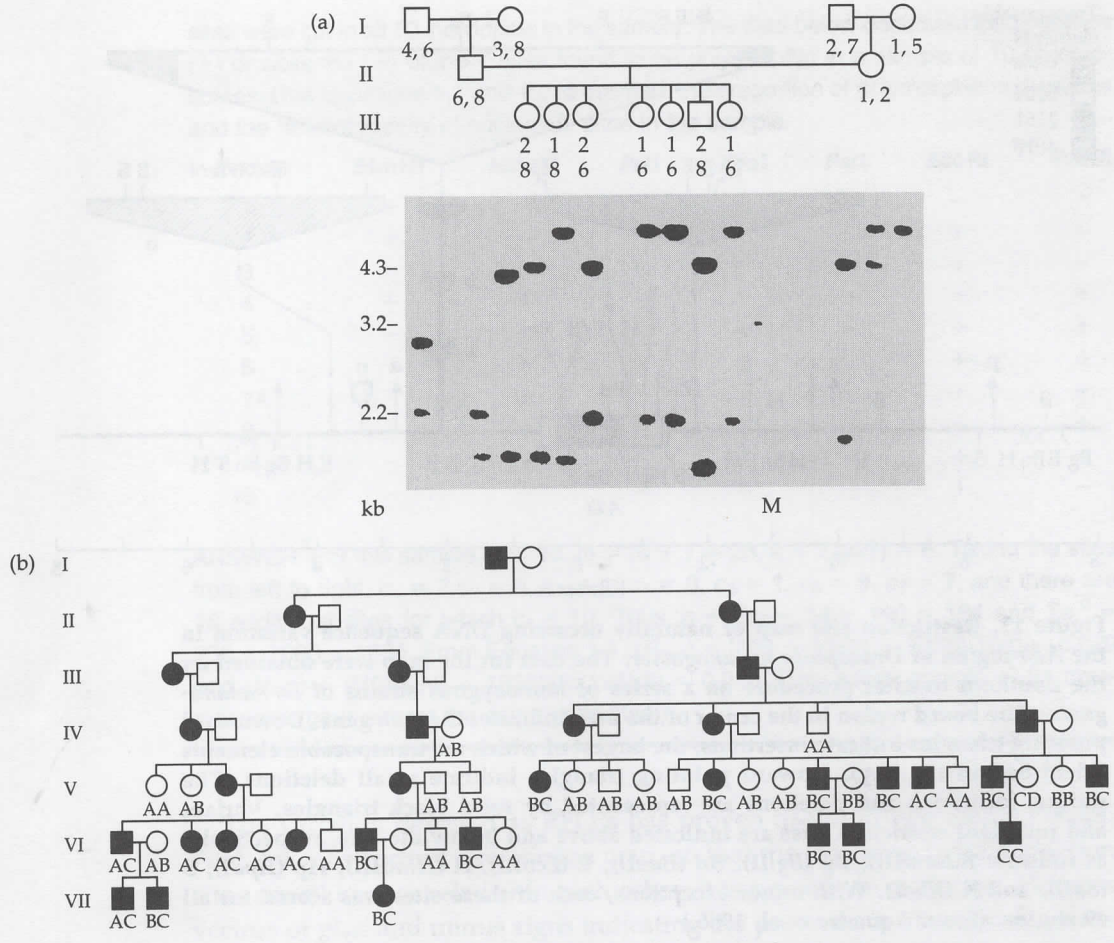


Figure 14. (a) Autoradiogram of a Southern transfer from a three-generation family showing Mendelian transmission of RFLPs. Panel A, DNA was digested with *Msp*I and probed with *pYNH24*; Panel B, DNA was digested with *Bam*HI and probed with *pYNZ22*. The numbers beneath each individual in the pedigree refer to allele symbols. (From Nakamura et al. 1987.) (b) A seven-generation pedigree of a Venezuelan family, in which members with Huntington disease (an autosomal dominant) are indicated by solid symbols. Letters beneath an individual indicate the RFLP type when DNA is probed with the *G8* probe. (After Gusella et al. 1983.)

ping (which shows that females produce a higher frequency of recombinant gametes than males), the human linkage map of the autosomes is approximately 90% longer in females than in males. The male map is not uniformly smaller than the female map, however; some regions exhibit higher rates of recombination in males than in females. This paradox raises many questions

about the evolution of the linkage map itself, a topic that we turn to in Chapter 4.

Population geneticists have focused a great deal of attention on the DNA found within the mitochondrion (mtDNA). Animal mitochondria have a closed circular DNA molecule about 17,000 base pairs in length, and it is highly conserved in gene content and order. However, the rate of sequence divergence of the mitochondrion in mammalian lineages is nearly ten times that of nuclear DNA. The occurrence of rapid sequence divergence in mtDNA has made it particularly useful for phylogenetic studies of closely related organisms. The mtDNA of an individual is relatively easy to purify, so sample sizes may be sufficiently large to obtain reasonable estimates of the important population genetic parameters. Figure 15 shows the variation detected among mtDNA in European populations of the house mouse, *Mus musculus*. We will explore some of the exciting findings from the studies of molecular sequence variation in Chapter 7.

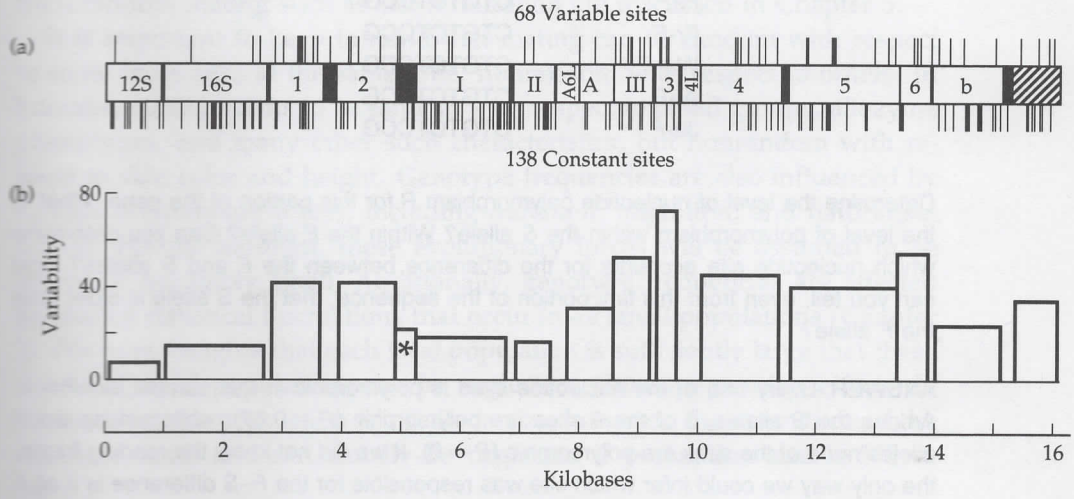


Figure 15. (a) Map of the variable and constant restriction sites in 38 types of *Mus domesticus* mitochondrial DNA detected with 11 restriction endonucleases. The lines above the horizontal bar showing the mtDNA genes indicate 68 variable restriction sites, and the lines below the bar indicate 138 constant restriction sites. The genes are indicated as follows: ribosomal genes (12S and 16S); tRNA genes or spacers (black areas); genes encoding cytochrome oxidases I, II, and III, ATPase, and cytochrome *b* (I, II, III, A, and b); NAD dehydrogenases (1-6, A6L, and 4L); and a displacement loop and adjacent noncoding regions (shaded portion). (b) The histogram shows the degree of variability in each gene; the variable sites are divided by the total sites and converted to percent. The bar with the asterisk indicates the variability for all 22 tRNAs pooled. (From Ferris et al. 1983a.)

E Allelic Variation in DNA Sequence

There has been considerable interest in the polymorphism of the alcohol dehydrogenase gene (*Adh*) in *Drosophila melanogaster* because it shows a polymorphism for a fast (Adh^F) and a slow (Adh^S) allele that is nearly worldwide in distribution. The alleles show parallel clines in frequency on several continents, and these data are often examined to evaluate different population genetic forces that might be influencing the polymorphism. Kreitman (1983) sampled 11 copies of the *Adh* locus from 6 slow alleles and 5 fast alleles. He sequenced 2721 bp in each of 11 clones, and a portion of the data (taken from the translated region of the fourth exon) are given below.

Strain	Haplotype
Wa-*S*	ACCCACTAG
Fl-*1S*	ACCCACTAG
Af-*S*	CTCCACTAA
Fr-*S*	CTCCACTAA
Fl-*2S*	CTCCACTAG
Ja-*S*	CTCTATTCA
Fl-*F*	CTGTCTCCG
Fr-*F*	CTGTCTCCG
Wa-*F*	CTGTCTCCG
Af-*F*	CTGTCTCCG
Ja-*F*	CTGTCTCCG

Determine the level of nucleotide polymorphism P for this portion of the gene. What is the level of polymorphism within the *S* allele? Within the *F* allele? Can you determine which nucleotide site accounts for the difference between the *F* and *S* alleles? How can you tell, even from this tiny portion of the sequence, that the *S* allele is older than the *F* allele?

ANSWER Every one of the nucleotide sites is polymorphic in this sample, so $P = 1$. Among the *S* alleles, 6 of the 9 sites are polymorphic ($P = 0.67$), while among the *F* alleles none of the sites are polymorphic ($P = 0$). If we did not know the reading frame, the only way we could infer which site was responsible for the *F*–*S* difference is if all *F* sequences had one nucleotide at a site, and all *S* sequences had another nucleotide at that site, but even this would not be definitive (both sites 3 and 5 satisfy this criterion). With knowledge of the reading frame, it becomes clear that position 5 accounts for the change from threonine in *F* to leucine in *S*. A newly arisen allele would not have had much time to accumulate silent polymorphisms, whereas an old allele could actually be a heterogeneous mixture of silent polymorphisms. By this reasoning, the *S* allele appears older than the *F* allele.

THE HARDY-WEINBERG PRINCIPLE

Random Mating and Random Union of Gametes

Predicting genotype frequencies from knowledge of allele frequencies is quite straightforward, but there are a few complications. Genotype frequencies are determined in part by mating patterns. (For example, plant species that regularly undergo self-fertilization are expected to have very low levels of heterozygosity.) One of the simplest and most important mating patterns is **random mating**, in which mating takes place at random with respect to the gene under consideration. With random mating, the chance that an individual mates with another having a prescribed genotype is equal to the frequency of that genotype in the population. For example, suppose that in some population a gene has genotypes *AA*, *Aa*, and *aa* in the proportions 0.16, 0.48, and 0.36, respectively; if mating is random, *AA* males will mate with *AA*, *Aa*, and *aa* females in the proportions 0.16, 0.48, and 0.36, respectively, and these same proportions will apply to the mates of *Aa* and *aa* males. In this section, we focus on random mating; effects of departures from random mating such as self-fertilization are discussed in Chapter 5.

It is important to keep in mind that mating can be random with respect to some traits and, at the same time, nonrandom with respect to others. In humans, mating seems to be random with respect to blood groups, allozyme phenotypes, and many other such characteristics, but nonrandom with respect to skin color and height. Genotype frequencies are also influenced by various evolutionary forces, including mutation, migration, and natural selection. For the moment, these evolutionary forces will be assumed to be absent or negligibly small. Additionally, genotype frequencies are affected by chance statistical fluctuations that occur in all small populations (Chapter 3). For now, imagine that each local population is sufficiently large that these effects of small populations can be neglected. (Small-population effects do occur unless the population is infinite in size; however, the magnitude of the effects is small and can usually be neglected if population size is 500 or more.)

The assumptions made so far in developing the model for predicting genotype frequencies may be summarized as follows:

1. The organism in question is diploid.
2. Reproduction is sexual.
3. Generations are nonoverlapping.
4. Mating is random.
5. Population size is very large.

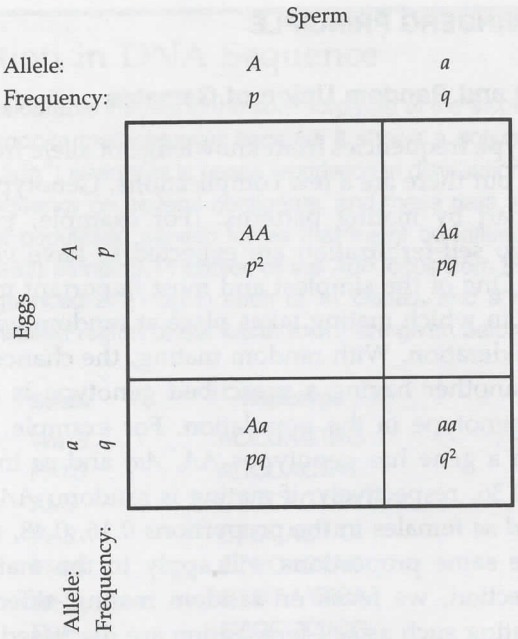


Figure 16. Punnett square showing Hardy-Weinberg frequencies generated by random mating with two alleles, *A* and *a*, which have frequencies *p* and *q*, respectively.

6. Migration is negligible.
7. Mutation can be ignored.
8. Natural selection does not affect the gene under consideration.

Collectively, these assumptions constitute the Hardy-Weinberg model, after G. H. Hardy (1908) and W. Weinberg (1908), who discovered how these assumptions result in a particular relation between gene and genotype frequencies. (See Li 1967a for a discussion of the largely overlooked early contributions of W. E. Castle.) Once the model is carefully stated, as in the preceding list, the genotype frequencies for a gene with two alleles can be worked out very easily because random mating of individuals is equivalent to random union of gametes. (You can verify this for yourself in Box F.) Because of this equivalence, we can cross multiply allele frequencies along the margins as in the Punnett square shown in Figure 16.

One Locus with Two Alleles

The result of the calculations in Figure 16 is the so-called **Hardy-Weinberg principle**: with random mating, the frequencies of the genotypes (usually called **genotype frequencies**) of *AA*, *Aa*, and *aa* are p^2, $2pq$, and q^2, respec-

tively, where p represents the allele frequency of A, q the allele frequency of a, and $p + q = 1$, since there are only two alleles. Cross multiplication in a Punnett square is simply a systematic way of going through all the possible gametic combinations. The probability that a sperm or egg carries A is p; the probability that a sperm or egg carries a is q. With random combination of gametes, the chance that an A-bearing sperm fertilizes an A-bearing egg is $p \times p = p^2$; therefore, this is the frequency of AA genotypes. The probability that an A-bearing sperm fertilizes an a-bearing egg is $p \times q = pq$; and the probability that an a-bearing sperm fertilizes an A-bearing egg is $q \times p = qp$. Altogether the frequency of Aa heterozygotes is $pq + qp = 2pq$. Finally, the probability that an a-bearing sperm fertilizes an a-bearing egg is $q \times q = q^2$, which is the genotype frequency of aa. Note that $p^2 + 2pq + q^2 = (p + q)^2 = (1)^2 = 1$, thus accounting for all the offspring. Graphs of p^2, $2pq$, and q^2 for various values of p and q are shown in Figure 17.

It is perhaps hard to believe that so simple a result from so simple a model

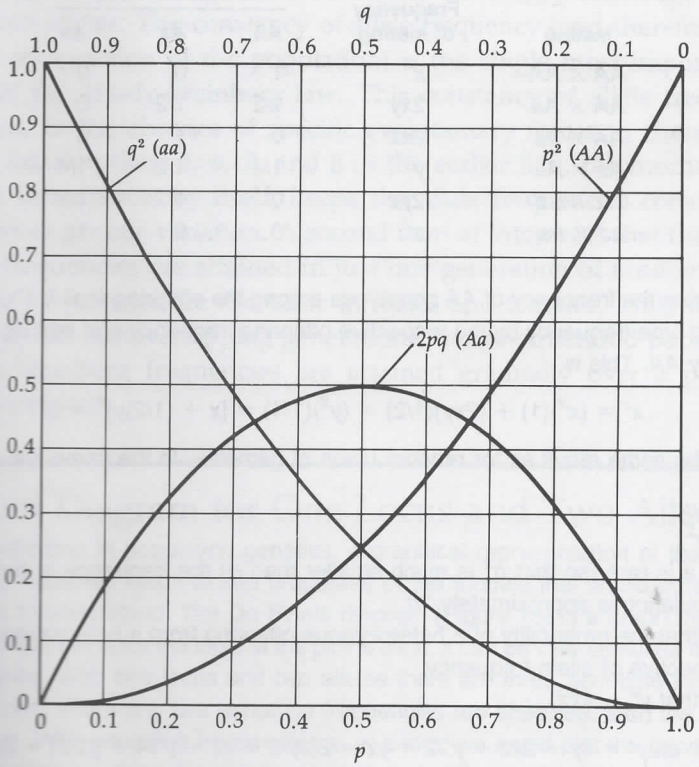


Figure 17. Graphs of p^2, $2pq$, and q^2. If the allele frequencies are between 1/3 and 2/3, heterozygotes will be the most common genotype in the population.

can hold so widely and be so important, but the Hardy-Weinberg principle provides the foundation for many investigations in population genetics. Several implications should be spelled out immediately. (Others are discussed in Box F.) First, the allele frequencies in the model population remain constant through time.

F Random Mating and Random Union of Gametes

In most cases, random mating of individuals is equivalent to random union of gametes. The Punnett square shows in a simple way how the Hardy-Weinberg principle is derived for random union of gametes. A bit more work is needed to show that random mating of individual genotypes yields the same result. Let the frequencies of *AA*, *Aa*, and *aa* genotypes be x, y, and z, where $x + y + z = 1$. The frequency of the *A* allele is $p = x + (1/2)y$ and that of the *a* allele is $q = (1/2)y + z$. The table below shows the frequencies of all possible matings, and the frequencies of offspring produced from each mating.

Mating	Frequency of mating	OFFSPRING FREQUENCIES		
		AA	*Aa*	*aa*
$AA \times AA$	x^2	1	0	0
$AA \times Aa$	$2xy$	1/2	1/2	0
$AA \times aa$	$2xz$	0	1	0
$Aa \times Aa$	y^2	1/4	1/2	1/4
$Aa \times aa$	$2yz$	0	1/2	1/2
$aa \times aa$	z^2	0	0	1

To determine the frequency of *AA* genotypes among the offspring (call it x'), we multiply the mating type frequency by the respective offspring frequency and add up the column headed by *AA*. This is

$$x' = (x^2)(1) + (2xy)(1/2) + (y^2)(1/4) = [x + (1/2)y]^2 = p^2$$

which is the same result as for random union of gametes. In the same way, show that:

a. $y' = 2pq$
b. $z' = q^2$
c. When *a* is rare (so that q^2 is much smaller than q) the frequency of heterozygotes in the population is approximately $2q$.
d. Show that the probability of a heterozygous offspring from a heterozygous parent is 1/2, irrespective of allele frequency.
e. Show that $y^2 = 4xz$.

ANSWER **a.** $y' = xy + 2xz + y^2/2 + yz = 2[xy/2 + xz + y^2/4 + yz/2] = 2(x + y/2)(y/2 + z) = 2pq$.
b. $z' = y^2/4 + yz + z^2 = (y/2 + z)^2 = q^2$.

c. The frequency of heterozygotes = $2pq = 2(1 - q)q = 2q - 2q^2 \approx 2q$ because q^2 is so much smaller than q.
d. The matings $Aa \times AA$, $Aa \times Aa$ and $Aa \times aa$ each produce one-half heterozygous offspring, so one-half the offspring from a heterozygous mother are heterozygous.
e. $y^2 = (2pq)^2 = 4p^2q^2 = 4xz$.

From the rule in Table 1 for calculating allele frequencies, the allele frequency of A among the offspring works out to be two times p^2 (because each AA homozygote carries two A alleles) plus $2pq$ (because each Aa heterozygote carries one A allele) divided by two (because each individual carries two alleles at the locus). In symbols, the allele frequency of A among the offspring is

$$(2p^2 + 2pq)/2 = p^2 + pq = p(p + q) = p \times 1 = p \qquad \textbf{(1.3)}$$

Allele frequencies do not change from generation to generation; in any generation, therefore, the genotype frequencies will be p^2, $2pq$, and q^2 for AA, Aa, and aa, and these frequencies constitute the **Hardy-Weinberg equilibrium** for two alleles. The constancy of allele frequency (and therefore of the genotypic composition of the population) is the single most important implication of the Hardy-Weinberg law. This constancy of allele frequencies implies that, in the absence of specific evolutionary forces to change allele frequency (assumptions 5, 6, 7, and 8 in the earlier list), the mechanism of Mendelian inheritance, by itself, keeps the allele frequencies constant and thus preserves genetic variation. A second item of interest is that the Hardy-Weinberg frequencies are attained in just one generation of random mating if the allele frequencies are the same in males and females. This, however, is true only with nonoverlapping generations; in age-structured populations, the Hardy-Weinberg frequencies are attained gradually over a period of several generations.

De Finetti Diagram for One Locus and Two Alleles

G

For many problems in population genetics, a graphical representation of the relationships among variables helps to see properties of the models that would otherwise be quite difficult to understand. The De Finetti diagram (Figure 18) is a graph that initially seems contrived, but once the logic of the plot is clear, it can be very useful for illustrating many principles. With one locus and two alleles there are three genotype frequencies that sum to one, and if any two genotype frequencies are specified, then the third one is also known. With genotype frequencies x, y, and z, we could plot the genotype pair (x,y) on an ordinary rectangular coordinate system, but the valid points would be limited to within the triangle (0,0), (1,0), and (0,1), with a distortion of the areas. The De Finetti

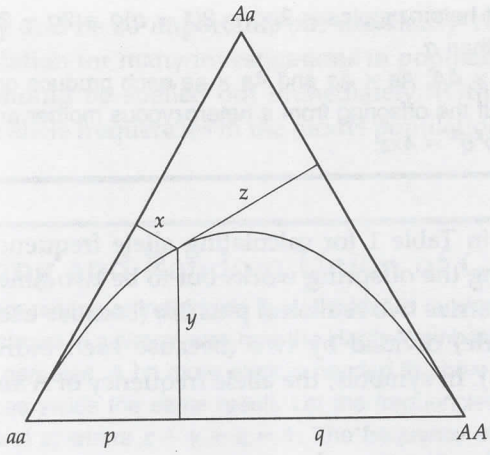


Figure 18. De Finetti diagram showing the frequencies of the three genotypes by location in a triangular coordinate system. The lengths of the line segments *x*, *y*, and *z* indicate the frequencies of genotypes *AA*, *Aa*, and *aa*, respectively. The vertices at the base represent fixation of the *a* allele (left) and the *A* allele (right). The top vertex represents a population with only heterozygotes and no homozygotes. The parabola that runs through the triangle is the set of points representing genotype frequencies in Hardy-Weinberg equilibrium.

diagram avoids these distortions by plotting the genotype frequencies within an equilateral triangle (Figure 18). The vertices represent points where the population has only one genotype. For example, as the labeling shows, the lower left corner has only *aa* individuals. To determine the genotype frequencies represented by any point in the triangle, measure the perpendicular distance from the side opposite the relevant vertex to the point, as indicated on the diagram. The frequency of heterozygotes is the height of the point above the base of the triangle. The line that is dropped down from the point to the base divides the base in the proportions p:q, thereby indicating the allele frequency.

It is important to note here that the Hardy-Weinberg principle is not very sensitive to certain kinds of departures from the assumptions numbered 3 to 8 in the earlier list, particularly numbers 5 to 8 (i.e., very large population size with no migration, mutation, or selection). Observed genotype frequencies in a population may fit the expected Hardy-Weinberg proportions; however, this cannot be taken as evidence that the assumptions in the list are valid. To be concrete, suppose that *AA*, *Aa*, and *aa* zygotes survive to adulthood in the proportions w_{11}, w_{12}, and w_{22}, respectively, but that the genotypes are otherwise completely equivalent and that assumptions 1 to 7 in the earlier list are satisfied. In such a case, *AA*, *Aa*, and *aa* adults will occur in Hardy-Weinberg proportions whenever $w_{12} = \sqrt{w_{11}w_{22}}$, whatever that actual values of w_{11}, w_{12}, w_{22} may be; thus, for example, the Hardy-Weinberg proportions

will hold even if $w_{11} = 1.0$, $w_{12} = 0.50$, and $w_{22} = 0.25$ — that is, even if three-fourths of the *aa* zygotes and one-half of the *Aa* zygotes die before reaching adulthood! If w_{12} does not equal $\sqrt{w_{11}w_{22}}$, there will be some deviation from Hardy-Weinberg proportions, but for plausible values of w_{11}, w_{12}, and w_{22} the deviations are expected to be small and detection therefore requires very large sample sizes (Lewontin and Cockerham 1959).

The Hardy-Weinberg Principle in Operation

Use of the Hardy-Weinberg principle is exemplified in Table 2, which gives the results of MN blood typing of 6129 American Caucasians: 1787 were genotypically MM, 3037 were MN, and 1305 were NN (Mourant et al. 1976). The allele frequencies of *M* and *N* are calculated and the Hardy-Weinberg frequencies obtained as shown. The expected number of each genotype is obtained by multiplying the Hardy-Weinberg frequencies by the sample size (6129). The goodness-of-fit between the observed numbers and the expected numbers can be judged quantitatively by means of the chi-squared value (χ^2), calculated as follows:

$$\chi^2 = \Sigma \frac{(\text{observed number} - \text{expected number})^2}{(\text{expected number})} \qquad \textbf{(1.4)}$$

with the sum taken of all genotypic classes, producing $\chi^2 = 0.05$. Purist

Table 2. Fit of Hardy-Weinberg frequencies to observed data.

	GENOTYPE			
	M/M	*M/N*	*N/N*	**Total**
Number of individuals	1787	3037	1305	6129
Number of *M* alleles	3574	3037	0	6611
Number of *N* alleles	0	3037	2610	5647
Number of *M* + *N* alleles	3574	6074	2610	12258
	Allele frequency of $M = 6611/12{,}258 = 0.53932 = p$			
	Allele frequency of $N = 5647/12{,}258 = 0.46068 = q$			
Expected frequency	$p^2 = 0.29087$	$2pq = 0.49691$	$q^2 = 0.21222$	1.000
Expected no. (frequency × 6129)	1782.7	3045.6	1300.7	6129
χ^2 value	$\frac{(1787 - 1782.7)^2}{1782.7} + \frac{(3037 - 3045.6)^2}{3045.6} + \frac{(1305 - 1300.7)^2}{1300.7} = 0.04887$			
	Degrees of freedom $= 3 - 1 - 1 = 1$			
	Probability associated with χ^2_1 of 0.04887 is about 0.90			

statisticians distinguish between the χ^2 distribution and the number calculated from Equation 1.4, the latter being called a test statistic, often written as X^2. The test based on this statistic is still called the "chi-square test."

For goodness-of-fit tests using χ^2 values, each test has an associated number called the **degrees of freedom**. The number of degrees of freedom is calculated as the total number of classes of data (in this case 3) minus the number of parameters estimated from the data (in this case 1, because p was estimated) minus 1. For the χ^2 test in Table 2, therefore, the number of degrees of freedom is $3 - 1 - 1 = 1$.

The best way to assess the goodness-of-fit of an observed set of data to an expected result is to calculate the probability of getting a fit as bad or worse than the observed data by chance. The null hypothesis is that the data perfectly fit the Hardy-Weinberg proportions, and in this case the χ^2 value would be 0. The greater the departure from Hardy-Weinberg proportions, the worse the fit and the higher the χ^2 value. The probability of getting a χ^2 value greater than the observed value by chance depends on the number of degrees of freedom, and in practice is either read from a table, or calculated by a rather cumbersome algorithm (Press et al. 1986). If the probability value for a particular χ^2 is P, this means that if we were to take 1000 samples of the same size from a population in Hardy-Weinberg equilibrium and calculate a χ^2 for each, a fraction P would be greater than our observed test value. Typically, if $P < 0.05$, the test is considered "significant," which means the null hypothesis is rejected, and the test indicates significant departure from Hardy-Weinberg proportions.

Returning now to Table 2, the probability value associated with a χ^2 of 0.04887 with 1 degree of freedom is about 0.90—a remarkably good fit. A χ^2 as large or larger would be obtained by chance 90% of the time in a population having its actual genotype frequencies as calculated from the Hardy-Weinberg law; thus there is no reason for thinking this population does not obey the Hardy-Weinberg law for the *MN* locus.

The MN blood groups are easy to work with in calculating allele frequencies because each genotype has a unique phenotype. However, in cases where one of two alleles is dominant, only two phenotypic classes can be recognized, the dominant type and the recessive type. One example is an antigen on the surface of red blood cells coded by the *D* allele in the Rh blood group system in humans. Genotypes *DD* and *Dd* are Rh^+ (positive), whereas *dd* is Rh^- (negative). Thus, *D* is dominant to *d*. In one test of 22,133 American Caucasians, 85.8% were found to be Rh^+ and 14.2% were Rh^- (Mourant et al. 1976). Letting p stand for the allele frequency of *D* and q stand for the allele frequency of *d*, the allele frequencies can be calculated by noting that, with random mating, the frequency of homozygous recessives, in this case 0.142, must equal q^2. So $q^2 = 0.142$ or $q = \sqrt{0.142} = 0.3768$. Then

$p = 1 - q = 1 - 0.3768 = 0.6232$, and the frequencies of *DD*, *Dd*, and *dd* are expected to be $p^2 = (0.6232)^2 = 0.3884$, $2pq = 2 \times 0.6232 \times 0.3768 = 0.4696$, and $q^2 = (0.3768)^2 = 0.1420$, respectively. These values could have been approximated by reading off the curves in Figure 17. (Note from the discussion of degrees of freedom in the previous section that there is no chance for a goodness-of-fit test here because, with only two phenotypic classes, both degrees of freedom are lost, one because of fixed sample size and the other because q was estimated from the data.)

The above sort of estimation permits calculation of the frequency of heterozygous carriers of recessive alleles. If the observed frequency of recessive homozygotes is R, then with random mating $\langle q \rangle = \sqrt{R}$ and the frequency of heterozygotes is estimated to be $2 \times (1 - \sqrt{R}) \times \sqrt{R}$. (As for $\langle H \rangle$ and $\langle P \rangle$, the notation $\langle q \rangle$ indicates a statistical estimator of q.) For cystic fibrosis — a severe disease involving abnormal glandular secretions due to an autosomal recessive — in Caucasians, $R = 0.0006$ (approximately; Newcombe 1964). The frequency of heterozygotes is therefore about $2 \times (1 - \sqrt{0.0006}) \times \sqrt{0.0006} = 2 \times 0.9755 \times 0.0245 = 0.0478$. Whereas only about 1 individual in 1667 is actually affected with cystic fibrosis (i.e., 1/0.0006), about 1 in 21 is a heterozygous carrier (i.e., 1/0.0478). This illustrates the important principle that the vast majority of rare recessive alleles are present in phenotypically normal heterozygotes and are therefore concealed.

The Hardy-Weinberg Principle with Three or More Alleles

Genotypic frequencies under random mating for genes with three alleles are shown in Figure 19. The alleles are labeled as A_1, A_2, and A_3 and the corre-

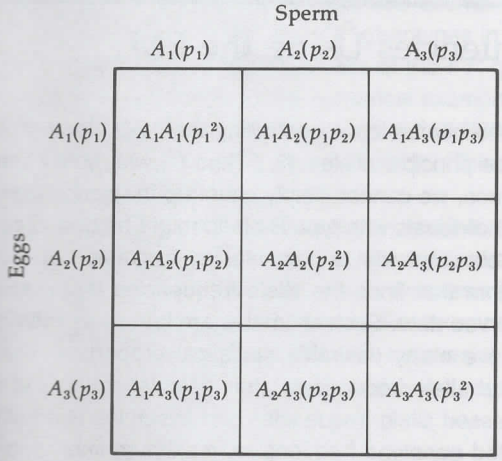


Figure 19. Punnett square showing Hardy-Weinberg frequencies for three autosomal alleles.

Table 3. Phenotype frequencies for ABO blood groups.

Genotype	Expected frequency	Blood type	Expected frequency	Observed frequency
I^O/I^O	$p_1^2 = (0.6708)^2 = 0.4500$	O	0.4500	0.4493
I^O/I^A	$2p_1p_2$ 2 (0.6708)(0.2703) = 0.3626	A	0.4357	0.4365
I^A/I^A	$p_2^2 = (0.2703)^2 = 0.0731$	A		
I^O/I^B	$2p_1p_3 = 2$ (0.6708)(0.0588) = 0.0789	B	0.0824	0.0832
I^B/I^B	$p_2^2 = (0.0588)^2 = 0.0035$	B		
I^A/I^B	$2p_2p_3 = 2$ (0.2703)(0.0588) = 0.0318	AB	0.0318	0.0309

sponding allele frequencies are p_1, p_2, and p_3. Since there are only three alleles, $p_1 + p_2 + p_3 = 1$. The genotype frequencies of A_1A_1, A_1A_2, A_2A_2, A_1A_3, A_2A_3, and A_3A_3 are p_1^2, $2p_1p_2$, p_2^2, $2p_1p_3$, $2p_2p_3$, and p_3^2, respectively, which can be obtained by expanding $(p_1A_1 + p_2A_2 + p_3A_3)^2$—a calculation that the Punnett square does automatically. In one test of 3977 Swiss from Zurich (Mourant et al. 1976) the allele frequencies of I^O, I^A, and I^B, which control the ABO blood groups, were estimated from the method in Box H to be $p_1 = 0.6708$ (for I^O), $p_2 = 0.2703$ (for I^A), and $p_3 = 0.0588$ (for I^B). With random mating, the genotype frequencies can be calculated according to Figure 19, leading to the results shown in Table 3. I^A and I^B are both dominant to I^O, so the expected frequency of blood type A is the sum of the genotype frequencies of I^O/I^A and I^A/I^A, and similarly for blood type B. As can be seen, the expected and the observed frequencies are in very close agreement.

H Estimating Allele Frequencies Using the EM Algorithm

Estimation of allele frequencies at the human locus controlling the ABO blood group is troublesome because there are three principal alleles, I^A, I^B and I^O, with both I^A and I^B dominant to I^O. Because of dominance, we cannot simply count up the genes from the observed genotypes; for example, individuals with type B blood might be genotype I^BI^B or I^BI^O. Allele frequencies can be obtained with a computer by a simple but elegant algorithm known as the EM algorithm that finds the allele frequencies that maximize the probability of obtaining the observed data. Such estimates are known as **maximum likelihood estimates**, and they have many desirable statistical properties. The EM algorithm works by first guessing the allele frequencies, then calculating the expected genotype frequencies using the guessed allele frequencies and assuming Hardy-Weinberg proportions. Using the expected genotype frequencies, maximum likelihood estimates of allele frequencies are obtained, and the cycle is repeated until allele frequencies converge.

Estimates of allele frequency can be calculated using the EM algorithm as follows: Let p, q and r denote the allele frequencies of I^A, I^B and I^O, and let freq(O), freq(A), freq(B) and freq(AB) denote the frequencies of the O, A, B, and AB blood group phenotypes. We get preliminary estimates of the allele frequencies based on the assumptions that the population is in Hardy-Weinberg equilibrium:

$$\langle r \rangle = \sqrt{\text{freq(O)}}$$

$$\langle p \rangle = 1 - \sqrt{\text{freq(B + O)}}$$

$$\langle q \rangle = 1 - \sqrt{\text{freq(A + O)}}$$

Usually $\langle p \rangle + \langle q \rangle + \langle r \rangle$ will not equal 1, but this can be corrected by dividing each by the sum. To see how the method works, let us consider a numerical example. Among 2060 Croatians, the following number of individuals of each phenotype were found (Mourant et al. 1976): 702 O, 862 A, 365 B, and 131 AB. From the above equations, we get $\langle r \rangle = \sqrt{[702/(702 + 862 + 365 + 131)]} = 0.5838$, $\langle p \rangle = 0.2803$, and $\langle q \rangle = 0.1287$. When these are normalized, we get $\langle r \rangle = 0.5880$, $\langle p \rangle = 0.2824$, and $\langle q \rangle = 0.1296$.

With these crude estimates of allele frequencies, the expected genotype counts (assuming Hardy-Weinberg equilibrium) are gotten as follows:

$$\begin{array}{lll} I^AI^A & x_1 & = \text{freq(A)}(p^2/(p^2+2pr)) \\ I^AI^O & x_2 & = \text{freq(A)}(2pr/(p^2+2pr)) \\ I^BI^B & x_3 & = \text{freq(B)}(q^2/(q^2+2qr)) \\ I^BI^O & x_4 & = \text{freq(B)}(2qr/(q^2+2qr)) \\ I^OI^O & x_5 & = \text{freq(O)} \\ I^AI^B & x_6 & = \text{freq(AB)} \end{array}$$

Notice that the I^OI^O and I^AI^B genotypes are unambiguously scored, so their estimated genotype counts are simply the observed counts. Freq(A) is the count of the A phenotype, which consists of I^AI^A and I^AI^O genotypes in the proportion $p^2 : 2pr$, so a fraction $p^2/(p^2 + 2pr)$ of the A phenotypes are of the I^AI^A genotype. When the crude estimates of the allele frequencies from the numerical example are substituted into these formulae, the estimated genotype counts are $\langle x_1 \rangle = 166.8865$, $\langle x_2 \rangle = 695.1135$, $\langle x_3 \rangle = 36.2316$, $\langle x_4 \rangle = 328.7684$, $\langle x_5 \rangle = 702$, and $\langle x_6 \rangle = 131$. This constitutes the E step, and we have the *E*xpected genotype frequencies.

The M step of the EM algorithm is to calculate the *M*aximum likelihood estimates of the allele frequencies *assuming* that the estimated genotype frequencies are the true genotype frequencies. This is done by a process called "gene counting," which consists of quite literally counting up the genes. For example, the count of the I^A alleles is twice the number of I^AI^A genotypes plus the number of I^AI^O genotypes plus the number of I^AI^B genotypes. To get frequencies, these counts are divided by the total number of genes in the sample:

$$\langle p \rangle = \frac{2x_1 + x_2 + x_6}{2(\Sigma x_i)} \quad \langle q \rangle = \frac{2x_3 + x_4 + x_6}{2(\Sigma x_i)} \quad \langle r \rangle = \frac{2x_5 + x_2 + x_4}{2(\Sigma x_i)}$$

In the numerical example, the estimates are $\langle p \rangle = 0.2815$, $\langle q \rangle = 0.1292$, and $\langle r \rangle = 0.5893$ after this step. This completes the M step. To obtain the maximum likelihood estimates of allele frequencies given the phenotype frequencies, we now return to the E step, estimate expected genotype frequencies, then perform the M step once again, and so on. This iterative process converges to stable values, which are the maximum likelihood values. In this case, convergence occurs after just five cycles. These values are $\langle p \rangle = 0.2813$, $\langle q \rangle = 0.1291$, and $\langle r \rangle = 0.5895$. The EM algorithm is easy to program into a computer, and is especially powerful when many parameters must be estimated simultaneously. For this reason, it has been used successfully in mapping 403 polymorphic genes in the human genome (Donis-Keller et al. 1987; Lander and Green 1987). For a detailed description of the EM algorithm, see Dempster et al. (1977).

In general, if there are n alleles $A_1, A_2, \ldots, A_n$ with frequencies $p_1, p_2, \ldots, p_n$ (where $\Sigma p_i = 1$) then the genotype frequencies under random mating are p_i^2 for A_iA_i homozygotes and $2p_ip_j$ for A_iA_j heterozygotes. The total frequency of heterozygotes is 1 minus the sum of the frequency of homozygotes, or $H = 1 - \Sigma p_i^2$. The maximum value for H occurs when all the alleles have equal frequency, or $p_i = 1/n$ for all n alleles. When all alleles have equal frequency, then $H = 1 - 1/n$, which shows that H increases with allelic diversity. Another way to see the effect of three alleles on heterozygosity is depicted in Figure 20. The contours show that whenever a new allele is introduced into a population that only has one or two alleles, H increases. This is intuitively clear because when a new allele is introduced, it is initially rare, so most copies of the allele are found in heterozygotes.

As an example of examining multiple alleles in a natural population, Taylor and Powell (1977) carried out microscopic examination of the karyotypes (chromosomal constitutions) of *Drosophila persimilis* larvae from Mather, California. Four different gene arrangements of the third chromosome were observed, the different gene arrangements being due to the presence of paracentric inversions. In one sample of 168 larvae, Taylor and Powell observed a frequency of 0.66 for the Whitney (*WH*) arrangement, 0.13 for Klamath (*KL*), 0.14 for Standard (*ST*), and 0.07 for Mendocino (*MD*). With four gene arrangements, there are four possible homokaryotypes (e.g., *WH/WH*) and six heterokaryotypes (e.g., *WH/KL*). In a random mating population, the frequency of any homokaryotype is expected to be the square of the corresponding frequency [i.e., *WH/WH* would have frequency $(0.66)^2 = 0.4356$]; the frequency of any heterokaryotype is expected to be twice the product of the corresponding chromosomal frequencies (i.e., *WH/KL* would have frequency $2 \times 0.66 \times 0.13 = 0.1716$). Expected frequencies of all ten karyotypes can be obtained by multiplying out the expression $(0.66WH + 0.13KL + 0.14ST + 0.07MD)^2$.

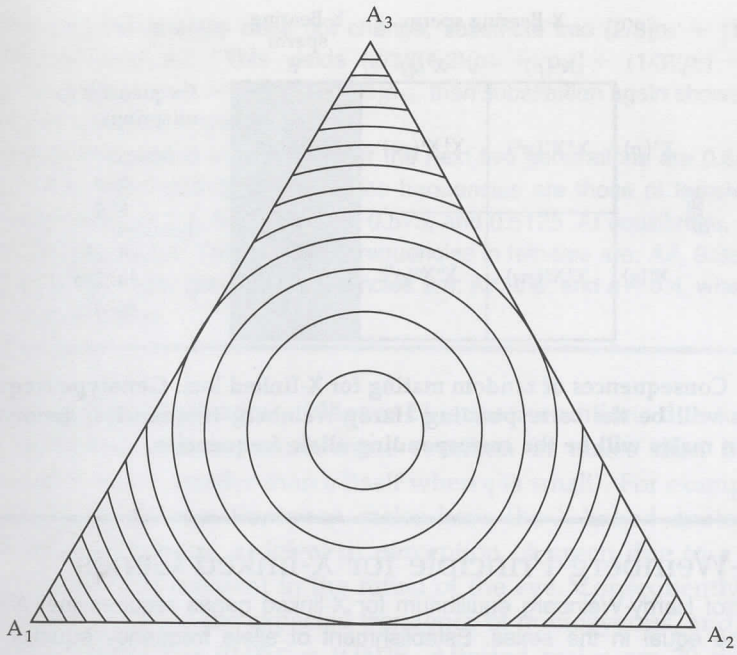


Figure 20. Contour plot of total heterozygosity, $H = 1 - \Sigma p_i^2$ on triangular (De Finetti) coordinates. The coordinates are slightly different from those in Figure 18, in that points in the triangle represent the frequencies of the three alleles A_1, A_2, and A_3. The edges of the triangle represent populations with just two alleles segregating, and the vertices represent fixations for a single allele. The center is the maximum, with $H = 2/3$. The interval between contours is 0.05.

The Hardy-Weinberg Principle for X-linked Genes

For genes on the human X chromosome, genotype frequencies differ in the sexes: females have two X chromosomes, and the genotype frequencies are as given by the Hardy-Weinberg principle; males have only one X chromosome, and the genotype frequencies are simply equal to the allele frequencies. The rule for X-linked genes is shown in Figure 21 for two alleles, and the rule applies to all organisms in which the females are XX and the males XY. This formulation holds only if allele frequencies for the X-linked locus are the same in eggs and sperm, but it can be shown that such equality between the sexes is established quite rapidly, over a period of ten or so generations even in age-structured populations (Box I).

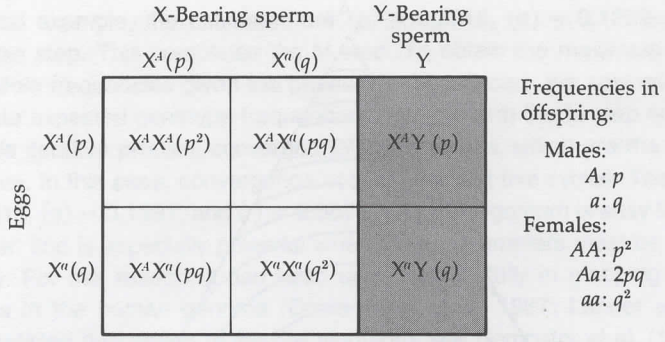


Figure 21. Consequences of random mating for X-linked loci. Genotype frequencies in females will be the corresponding Hardy-Weinberg frequencies; genotype frequencies in males will be the corresponding allele frequencies.

I Hardy-Weinberg Principle for X-linked Genes

Attainment of Hardy-Weinberg equilibrium for X-linked genes requires that allele frequencies be equal in the sexes. Establishment of allele frequency equality occurs gradually over several generations, and it is governed by two principles resulting from the mechanism of X-linked inheritance. If we consider an X-linked gene with alleles *A* and *a*, and let p_m and p_f be the frequency of allele *A* in males and females respectively, then the two principles are:

1. $p_m' = p_f$
2. $p_f' = (1/2)(p_m + p_f)$

The first principle states that the frequency of allele *A* in males is the same as the allele frequency in females of the previous generation, because males get their X chromosome exclusively from their mother. The second principle is that the female allele frequency is the average of the frequency in males and females of the previous generation, because females receive one X chromosome from each parent.

a. Prove that the difference in allele frequency between the sexes is halved each generation, and that the differences alternate in sign. This implies that

$$p_f' - p_m' = (-1/2)(p_f - p_m)$$

b. Prove that the quantity $(2/3)p_f + (1/3)p_m$ remains constant. Prove that if the initial allele frequency in both sexes equals this quantity, then the allele frequency does not change through time.

c. For initial frequencies $p_m = 0.2$ and $p_f = 0.8$, calculate the allele frequencies in the two sexes for the next five generations, and calculate the equilibrium allele and genotype frequencies.

ANSWER **a.** By substitution, $p_f' - p_m' = (1/2)(p_m + p_f) - p_f = (1/2)p_m - (1/2)p_f = -(1/2)(p_m + p_f)$.

b. To see that the quantity does not change, substitute into $(2/3)p_f' + (1/3)p_m'$ the values for p_f' and p_m'. This yields $(2/3)[(1/2)(p_m + p_f)] + (1/3)(p_f) = (2/3)p_f + (1/3)p_m$. If we let $p_f = p_m = (2/3)p_f + (1/3)p_m$, then substitution again shows that these quantities do not change over time.

c. The allele frequencies in females over the next five generations are 0.8, 0.5, 0.65, 0.575, 0.6125, and 0.60313. In males, the frequencies are those of females the preceding generation, or 0.2, 0.8, 0.5, 0.65, 0.575, and 0.6125. At equilibrium, $p_f = p_m = (2/3)p_f + (1/3)p_m = 0.6$. The genotype frequencies in females are: *AA*, 0.36; *Aa*, 0.48; and *aa*, 0.16. The male genotype frequencies are: *AY*, 0.6; and *aY*, 0.4, where *Y* refers to the Y chromosome.

The important implication of Figure 21 is that conditions due to rare sex-linked recessives will be much more common in males than in females (because q^2 is much smaller than q itself when q is small). For example, about five percent of Western European males have the X-linked deuteranomaly type of color blindness, a defect in perception of green due to a defective version of a certain pigment in the retina of the eye. Consequently $q = 0.05$ for this condition, and the expected frequency of homozygous, and therefore color blind, females is $(0.05)^2 = 0.0025$. Affected males are 20 times more frequent than affected females. In the case of the X-linked protanomaly type of color blindness, which involves a different pigment, the frequency of affected males is about 0.01 in Western Europeans. The expected frequency of protanomalous females due to this X-linked allele is therefore $(0.01)^2 = 0.0001$. In this case affected males are 100 times more frequent than affected females. These examples illustrate the principle that, for rare sex-linked recessives, the rarer the allele, the more the affected phenotype will tend to be found exclusively in males.

The Hardy-Weinberg Principle for Two Genes

With random mating, the alleles at any locus will rapidly acquire random association into genotypes in the frequencies shown in Figures 16 and 19. For two autosomal genes (autosomal genes are located in autosomes rather than in sex chromosomes), call them the A gene (with two alleles, A_1 and A_2, at frequencies denoted p_1 and p_2) and the B gene (also with two alleles, B_1 and B_2, at frequencies denoted q_1 and q_2), the genotype frequencies at the A locus will be p_1^2 for A_1A_1, $2p_1p_2$ for A_1A_2, and p_2^2 for A_2A_2; the genotype frequencies at the B locus will be q_1^2 for B_1B_1, $2q_1q_2$ for B_1B_2, and q_2^2 for B_2B_2. However, the alleles of one gene may not be in random association in gametes with the alleles of the other gene. Given time, the alleles of the genes will reach random association in gametes, but the approach is gradual and the rate of approach is slower the smaller the recombination fraction (r) between

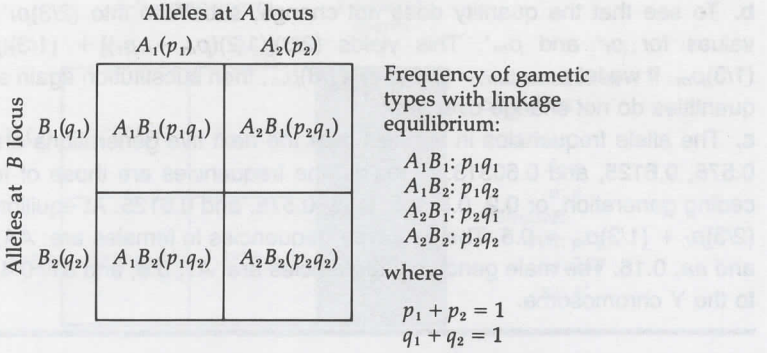


Figure 22. Random association between two alleles at each of two loci, showing expected gametic frequencies when the loci are in linkage equilibrium.

the genes. The meaning of "random association in the gametes" is shown in Figure 22. Here the combinations of *A*'s and *B*'s in the squares refer to *gametes*, not to genotypes, as in earlier figures. A state of random gametic association between alleles of different genes is called **linkage equilibrium**.

To be somewhat more precise about the effect that recombination has on the approach to linkage equilibrium, suppose that the actual frequencies of A_1B_1, A_1B_2, A_2B_1, and A_2B_2 gametes are P_{11}, P_{12}, P_{21}, and P_{22}, respectively where $P_{11} + P_{12} + P_{21} + P_{22} = 1$. (Sorry about the double subscripts here, but this is really the simplest sort of notation for two genes.) In these symbols, linkage equilibrium is defined as the state in which

$$\begin{aligned} P_{11} &= p_1q_1 \\ P_{12} &= p_1q_2 \\ P_{21} &= p_2q_1 \\ P_{22} &= p_2q_2 \end{aligned} \qquad (1.5)$$

We were able to express the frequencies of the three genotypes of one gene with two alleles in a two-dimensional triangle (De Finetti diagram), and a logical extension of this form of graph can be made to the model with two diallelic genes. In this case the four gamete types can be represented in three dimensions, because whenever we know three of the frequencies, the fourth one is determined. The simplest geometric form that spans a three-dimensional space is a tetrahedron, and points within a tetrahedron can represent the frequencies of the four gametic types (Figure 23). Cutting through the tetrahedron is a curved surface that indicates populations in linkage equilibrium. It is sometimes called the $D = 0$ surface; D will be defined shortly

(Equation 1.9). This geometric representation is especially helpful in understanding properties of the two locus model with selection, as we will see in Chapter 4.

Suppose for the moment that the genes are *not* in linkage equilibrium. (Genes that are not in linkage equilibrium are said to be in **linkage disequilibrium**.) We want to find out the gametic frequencies in the next generation. Consider first the A_1B_1 gamete. In any one generation, there are only two things that could have happened to a chromosome carrying A_1 and B_1: (1) it could have undergone recombination between the genes (an event that would occur with probability r, where r denotes the recombination fraction), or (2) the chromosome might have failed to undergo recombination between the genes (with probability $1 - r$). Among those chromosomes that did not undergo recombination, the frequency of A_1B_1 must be the same as it was in the previous generation; among the chromosomes that did undergo recombination, the frequency of A_1B_1 chromosomes must simply equal the fre-

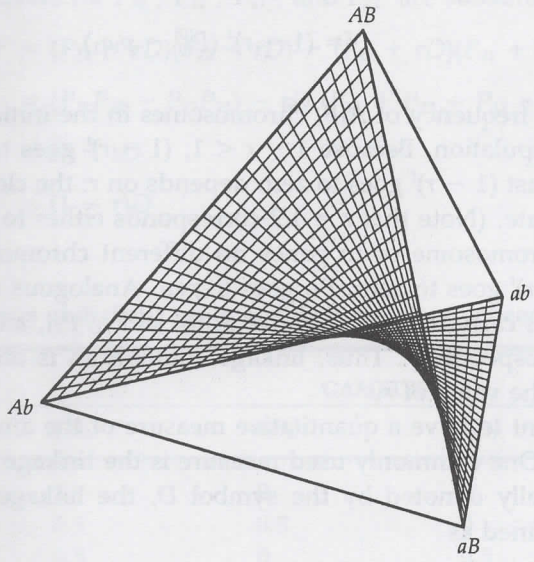


Figure 23. Tetrahedron representing two-locus gametic frequencies. The vertices represent fixation for the labeled gamete, and points in the interior have all four gametic types. Frequencies are represented by the distance from the opposite face to the labeled vertex. The center of the tetrahedron has all four gametes present in equal frequency. The curved surface cutting through the tetrahedron represents $D = 0$. Note that points along the edge from *Ab* to *AB* are on the $D = 0$ surface, but the points along the edge from *Ab* to *aB* are not on the surface. The *Ab*–*aB* and *AB*–*ab* edges have high values of linkage disequilibrium.

quency of A_1 ?/? B_1 genotypes in the previous generation (the question mark denotes any allele), and this is p_1q_1 because mating is random. Altogether, the frequency of A_1B_1 in the next generation, call it P_{11}', is

$$P_{11}' = (1 - r)P_{11} \text{ [nonrecombinants]} + rp_1q_1 \text{ [recombinants]} \quad \textbf{(1.6)}$$

Here P_{11} denotes the frequency of A_1B_1 in the previous generation. Subtraction of p_1q_1 from both sides leads to

$$P_{11}' - p_1q_1 = (1 - r)(P_{11} - p_1q_1) \quad \textbf{(1.7)}$$

This equation shows that the deviation of a gametic frequency from the product of its constituent allele frequencies decreases by a factor of $(1 - r)$ each generation. Now let us use the notation $P_{11}^{(t)}$ to indicate the frequency of the A_1B_1 gamete at generation t. The solution of this equation is as follows:

$$\begin{aligned} P_{11}^{(t)} - p_1q_1 &= (1 - r)(P_{11}^{(t-1)} - p_1q_1) \\ &= (1 - r)^2 (P_{11}^{(t-2)} - p_1q_1) \\ &= \ldots \\ &= (1 - r)^t (P_{11}^{(0)} - p_1q_1) \end{aligned} \quad \textbf{(1.8)}$$

where $P_{11}^{(0)}$ is the frequency of A_1B_1 chromosomes in the initial generation in the founding population. Because $1 - r < 1$, $(1 - r)^t$ goes to zero as t gets large, but how fast $(1 - r)^t$ goes to zero depends on r: the closer r is to zero, the slower the rate. (Note that $r = 1/2$ corresponds either to genes far apart on the same chromosome or to genes on different chromosomes.) In any case, since $(1 - r)^t$ goes to zero, P_{11} goes to p_1q_1. Analogous arguments hold for chromosomes carrying A_1B_2, A_2B_1, or A_2B_2, so P_{12}, P_{21}, and P_{22} go to p_1q_2, p_2q_1, and p_2q_2, respectively. Thus, linkage equilibrium is attained at a rate determined by the value of r.

It is convenient to have a quantitative measure of the amount of linkage disequilibrium. One commonly used measure is the **linkage disequilibrium parameter**. Usually denoted by the symbol D, the linkage disequilibrium parameter is defined as

$$D = P_{11}P_{22} - P_{12}P_{21} \quad \textbf{(1.9)}$$

Table 4 presents the ten different genotypes that are possible for two diallelic genes, and for each genotype it reports the frequencies of the gametes that they produce. Implicit in Table 4 is the assumption that males and females have an equal frequency of recombination. We can determine the frequency of the gamete A_1B_1 the next generation by multiplying the frequency of each genotype by the gametic frequency in the column headed by A_1B_1 and summing. This yields

$$
\begin{aligned}
P_{11}' &= P_{11}^2 + P_{11}P_{12} + P_{11}P_{21} + (1 - r)P_{11}P_{22} + rP_{12}P_{21} \\
&= P_{11}(P_{11} + P_{12} + P_{21} + P_{22}) - r(P_{11}P_{22} - P_{12}P_{21}) \\
&= P_{11} - rD
\end{aligned} \tag{1.10}
$$

Similarly, the other gametic frequencies change each generation according to the equations

$$
\begin{aligned}
P_{12}' &= P_{12} + rD \\
P_{21}' &= P_{21} + rD \\
P_{22}' &= P_{22} - rD
\end{aligned} \tag{1.11}
$$

Now, to get the recurrence equation that shows the way D changes over generations, note that

$$
D' = P_{11}'P_{22}' - P_{12}'P_{21}' \tag{1.12}
$$

When the equations for P_{11}', P_{12}', P_{21}', and P_{22}' are substituted, we get

$$
\begin{aligned}
D' &= (P_{11} - rD)(P_{22} - rD) - (P_{12} + rD)(P_{21} + rD) \\
&= (P_{11}P_{22} - P_{12}P_{21}) - rD(P_{11} + P_{22} + P_{12} + P_{21}) \\
&= D - rD \\
&= (1 - r)D
\end{aligned} \tag{1.13}
$$

Table 4. Genotypes of diallelic genes and their gametic frequencies.[a]

	GAMETES			
Genotype	A_1B_1	A_1B_2	A_2B_1	A_2B_2
A_1B_1/A_1B_1	1	0	0	0
A_1B_1/A_1B_2	0.5	0.5	0	0
A_1B_1/A_2B_1	0.5	0	0.5	0
A_1B_1/A_2B_2	$(1/2)(1 - r)$	$(1/2)r$	$(1/2)r$	$(1/2)(1 - r)$
A_1B_2/A_1B_2	0	1	0	0
A_1B_2/A_2B_1	$(1/2)r$	$(1/2)(1 - r)$	$(1/2)(1 - r)$	$(1/2)r$
A_1B_2/A_2B_2	0	0.5	0	0.5
A_2B_1/A_2B_1	0	0	1	0
A_2B_1/A_2B_2	0	0	0.5	0.5
A_2B_2/A_2B_2	0	0	0	1

[a] Assuming males and females have an equal frequency of recombination.

So each generation, the value of the linkage disequilibrium parameter is multiplied by a factor of $1 - r$. To extend this over many generations, the recursion becomes

$$D_t = (1 - r)^t D_0 \tag{1.14}$$

where D_t denotes the value of D in generation t. D therefore converges to a value of $D = 0$, at which point the population is said to be in linkage equilibrium. Figure 24 shows how D decays to zero for various values of r.

Another way to write the relations between gamete frequencies and allele frequencies is

$$\begin{aligned} P_{11} &= p_1 q_1 + D \\ P_{12} &= p_1 q_2 - D \\ P_{21} &= p_2 q_1 - D \\ P_{22} &= p_2 q_2 + D \end{aligned} \tag{1.15}$$

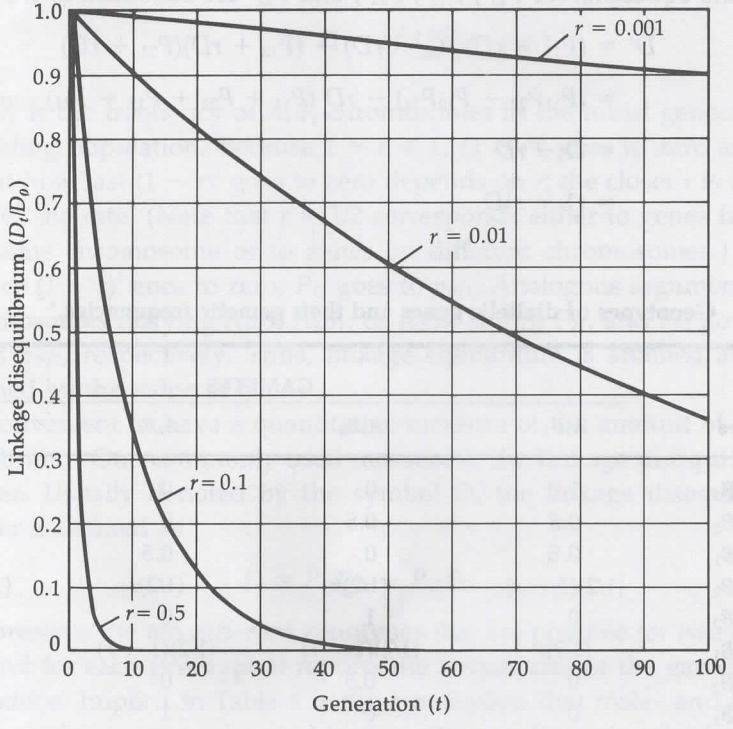


Figure 24. Decay of linkage disequilibrium parameter (D) with time, relative to its initial value (D_o), for various values of the recombination fraction between the loci (r). The graphs assume random mating. If there is substantial self-fertilization (as occurs in many plants), D decays much more slowly than shown here.

These equations show that D measures the departure of the gamete frequencies from the simple product of their component allele frequencies. The first equation in 1.15 can be rearranged as

$$D = P_{11} - p_1q_1 \tag{1.16}$$

Using Equation 1.16 as a definition of D, the equivalence of Equations 1.7 and 1.13 in deriving the change in D over time should be apparent.

Another useful measure of disequilibrium is

$$D' = D/D_{max} \tag{1.17}$$

where D_{max} is the maximum value that D can attain with the given allele frequencies (Lewontin 1964). Since $P_{12} = p_1q_2 - D \geq 0$, it follows that $D \leq p_1q_2$. Also, since $P_{21} = p_2q_1 - D \geq 0$, it follows that $D \leq p_2q_1$. Therefore, the maximum possible value of D (D_{max}) is given by

$$D_{max} = \min(p_1q_2, p_2q_1) \tag{1.18}$$

where "min" means "the smaller of." By similar reasoning, D is constrained to be greater than a minimum value found at

$$D_{min} = \max(-p_1q_1, -p_2q_2) \tag{1.19}$$

where "max" means "the larger of." Note that the largest possible value that D can have is 0.25 (when $P_{11} = P_{22} = 0.5$ and $P_{12} = P_{21} = 0$) and the smallest possible value is $D = -0.25$ (when $P_{12} = P_{21} = 0.5$ and $P_{11} = P_{22} = 0$). For both the maximal and minimal value of D, the allele frequencies are $p_1 = p_2 = q_1 = q_2 = 0.5$.

Values of D calculated for allozyme genes in natural populations are typically zero or very close to zero (indicating linkage equilibrium) unless the genes are very closely linked. Mukai et al. (1971, 1974) and Langley et al. (1977, 1978) found that there was very little linkage disequilibrium between genes that were more than ten map units apart, but for genes that were tightly linked there was a greater chance that they would be found in linkage disequilibrium. Highly significant values of D have been found in natural populations of plants (Clegg et al. 1972; Allard et al. 1972; Weir et al. 1972) and *Drosophila* (Baker 1975; Klarenberg and Scharloo 1986). On the other hand, at the nucleotide sequence level, the frequency of recombination is on the order of one percent per million base pairs in humans. In examining sequences of one or two thousand bases, the very low levels of recombination tend to result in very high levels of linkage disequilibrium, but there is surprising variation in the magnitude of linkage disequilibrium at the nucleotide level (see Chapter 7).

An extreme, naturally occurring example of linkage disequilibrium involves the paracentric inversions that are polymorphic in populations of

Drosophila pseudoobscura. The inversions prevent recombination; consequently, each inversion represents a sort of "supergene," and natural selection may accumulate beneficially interacting alleles (said to show **genetic coadaptation**) within each inversion (Prakash and Lewontin 1968; Zouros 1976; Prakash 1977; Voelker et al. 1978).

A case of linkage disequilibrium between an allozyme gene (*Amy* = α-amylase) and a paracentric inversion (*NS* = Nova Scotia) in *D. melanogaster* is provided by the data of Mukai et al. (1974) in Box B. Among the 1158 chromosomes examined, the observed numbers of the four chromosomal types were

Amy^F	non-*NS*	898
Amy^F	*NS*	122
Amy^S	non-*NS*	137
Amy^S	*NS*	1

If we denote the allele frequency of Amy^F as p_1 and that of Amy^S as p_2, then $\langle p_1 \rangle = (898 + 122)/1158 = 0.8808$ and $\langle p_2 \rangle = 1 - \langle p_1 \rangle = 0.1192$. Similarly, if the frequency of non-*NS* chromosomes is denoted q_1 and that of *NS* chromosomes as q_2, then $\langle q_1 \rangle = (898 + 137)/1158 = 0.8938$ and $\langle q_2 \rangle = 1 - \langle q_1 \rangle = 0.1062$. With linkage equilibrium, the corresponding expected numbers of each of the chromosome types would be

Amy^F	non-*NS*	$1158 \times 0.8808 \times 0.8938 = 911.65$
Amy^F	*NS*	$1158 \times 0.8808 \times 0.1062 = 108.32$
Amy^S	non-*NS*	$1158 \times 0.1192 \times 0.8938 = 123.37$
Amy^S	*NS*	$1158 \times 0.1192 \times 0.1062 = 14.66$

The χ^2 value for the comparison of the observed data with these expectations is 16.17; the number of degrees of freedom is 4 (the number of classes of data) minus 1 (for using the sample size in calculating expectations) minus 1 (for estimating p_1) minus 1 (for estimating q_1) or $4 - 1 - 1 - 1 = 1$. The probability level associated with a χ^2 of 16.17 with one degree of freedom is less than 0.0001, so we may conclude with considerable confidence that *Amy* and *NS* are not in linkage equilibrium.

As for the amount of linkage disequilibrium involved in the preceding example, let P_{11}, P_{12}, P_{21}, and P_{22} represent the frequencies of the four chromosome types, given in the same order as the observed numbers. Then the estimates are

$$\langle P_{11} \rangle = 898/1158 = 0.7755$$

$$\langle P_{12} \rangle = 122/1158 = 0.1054$$

$$\langle P_{21} \rangle = 137/1158 = 0.1183$$

$$\langle P_{22} \rangle = 1/1158 = 0.0009$$

and $\langle D \rangle = \langle P_{11} \rangle \langle P_{22} \rangle - \langle P_{12} \rangle \langle P_{21} \rangle = -0.0118$. Since $-p_1 q_1 = -0.7873$ is smaller than $-p_2 q_2 = -0.0127$, $D_{\min} = -0.0127$. The observed value of D is about 93% of its smallest possible value. Because the *Amy* gene is included within the *NS* inversion, exchange in "double heterozygotes" can only occur by double crossing over. This is a very rare event, so it would require a very long time for the linkage disequilibrium between *Amy* and *NS* to break down.

Linkage Disequilibrium and the Correlation Coefficient

J

The linkage disequilibrium parameter D measures the statistical association of alleles in forming gametes, and is actually related to the well known Pearson correlation coefficient (designated ρ, and defined below). That is, the parameter D is related to ρ to the extent that alleles of different genes are correlated in their occurrence in chromosomes. To see this, we assign values of 1 and 0 to random variables x and y, which represent effects of genes A and B respectively. We then calculate the correlation coefficient as follows:

Gamete	Frequency (f_i)	Value of x (A gene)	Value of y (B gene)	xy
A_1B_1	P_{11}	1	1	1
A_1B_2	P_{12}	1	0	0
A_2B_1	P_{21}	0	1	0
A_2B_2	P_{22}	0	0	0

The variance is defined as

$$\text{var}(x) = \overline{x^2} - \bar{x}^2$$

where $\overline{x^2} = \Sigma f_i x_i^2$ and $\bar{x} = \Sigma f_i x_i$ and the covariance is

$$\text{cov}(x,y) = \overline{xy} - \bar{x}\bar{y}$$

where $\overline{xy} = \Sigma f_i x_i y_i$. In this case we find

$$\bar{x} = P_{11} + P_{12} = p_1$$

$$\overline{x^2} = P_{11} + P_{12} = p_1$$

$$\bar{y} = P_{11} + P_{21} = q_1$$

$$\overline{y^2} = P_{11} + P_{21} = q_1$$

$$\overline{xy} = P_{11}$$

so after substituting, we obtain

$$\text{var}(x) = p_1 - p_1^2 = p_1 p_2$$

$$\text{var}(y) = q_1 - q_1^2 = q_1 q_2$$

$$\text{cov}(x,y) = P_{11} - p_1 q_1$$

The linkage disequilibrium parameter D is thus the covariance of allelic values. The definition of the Pearson correlation coefficient is

$$\rho = \frac{\text{cov}(x,y)}{\sqrt{\text{var}(x)\ \text{var}(y)}}$$

When the covariance and variances are substituted, we obtain

$$\rho = \frac{D}{\sqrt{p_1 p_2 q_1 q_2}}$$

Thus a positive value of D implies a positive correlation of alleles, implying an inflated chance of obtaining the A_1 and B_1 alleles (and the A_2 and B_2 alleles) together on a chromosome. A negative D implies a negative correlation, or a deficit of A_1B_1 and A_2B_2 chromosomes. Negative linkage disequilibrium also implies an excess of A_1B_2 and A_2B_1 chromosomes, again emphasizing that what we call positive and what we call negative linkage disequilibrium depends on the arbitrary labeling of alleles.

Two human examples showing linkage disequilibrium of closely linked genes are given in Table 5, where the data are drawn from the English population (Race and Sanger 1975). One case involves the MN blood group gene and a closely linked gene (alleles *S* and *s*) that controls related blood cell antigens. The other example involves two genes involved in the Rh blood group system. In the former case, D is about 48% of its maximum possible value; in the Rh case, D is about 81% of its hypothetical maximum. (For examples of linkage disequilibrium in humans between the HLA complex, which is involved in the immune response, and various genes resulting in increased susceptibility to such diseases as ankylosing spondylitis, see Thomson and Bodmer 1977; Hedrick et al. 1986.)

Linkage disequilibrium such as that seen in the preceding examples can be caused by linkage disequilibrium in the founding population that has not yet had time to dissipate due to the small value of r. Another possible cause is admixture of populations with differing gametic frequencies. A third possibility is the occurrence of a sufficient intensity of natural selection in favor of certain heterozygous genotypes to overcome the natural tendency for D to go to zero. A fourth possibility for generating linkage disequilibrium is that, purely by chance, some chromosomes may "drift" to high frequency. All of these mechanisms will be explored in detail in later chapters.

Table 5. Linkage disequilibrium for two human blood group systems.

System	Gamete	Frequency	Allele	Frequency	D	D_{max}	$D/D_{max} \times 100$ (percent)
MNSs	*MS*	0.247	*M*	0.530	0.0737	0.1537	47.9
	Ms	0.283	*N*	0.470			
	NS	0.080	*S*	0.327			
	Ns	0.390	*s*	0.673			
Rh	*DE*	0.144	*D*	0.590	0.0520	0.0640	81.2
	De	0.446	*d*	0.410			
	dE	0.012	*E*	0.156			
	de	0.398	*e*	0.844			

ESTIMATING LINKAGE DISEQUILIBRIUM

Where a population sample is assayed for the phenotypes of two codominant genes, there are a total of nine two-locus phenotypes. The two double heterozygotes cannot be distinguished without doing progeny testing (examining progeny from a test cross), so we cannot directly count up the frequencies of the four gamete types. The above example from Mukai et al. (1974) yielded direct counts of gametes because second chromosomes of *Drosophila melanogaster* were isolated or extracted by the use of special chromosomes. This entails a considerable cost in terms of labor and effort, and is not possible with most organisms. Without these direct counts, we cannot directly determine the linkage disequilibrium among the loci from such a sample. One partial exception to this concerns X-linked genes, since males display an X-linked genotype that represents maternal gametes. There is, however, a powerful means of obtaining a maximum likelihood estimate of the gamete frequencies, and hence of linkage disequilibrium, using the method of gene counting (Hill 1974). This method consists of getting the best fit between the observed and expected phenotype counts based on the genetic model that relates the relative counts of the genotypes (see Box K). One important statistical consideration is the *power* of the test of linkage disequilibrium. This is the probability of detecting $D \neq 0$, given that a pair of genes actually are in linkage disequilibrium. A very small sample would fail to prove significant linkage disequilibrium, and Thompson et al. (1988) show that often dauntingly large sample sizes are needed to detect significant values of D. Additional complications to the problem of estimating linkage disequilibrium are explored in Weir and Cockerham (1978, 1979) and Cockerham and Weir (1977).

K Estimating Linkage Disequilibrium for Two Diallelic Genes

A sample of two diallelic codominant genes yields nine classes of phenotypes which can be arrayed in a table as follows:

	B_1B_1	B_1B_2	B_2B_2
A_1A_1	N_{11}	N_{12}	N_{13}
A_1A_2	N_{21}	N_{22}	N_{23}
A_2A_2	N_{31}	N_{32}	N_{33}

These N's are the observed numbers of the nine genotypes. The *expected* frequencies are obtained from the products of the gamete frequencies:

	B_1B_1	B_1B_2	B_2B_2
A_1A_1	P_{11}^2	$2P_{11}P_{12}$	P_{12}^2
A_1A_2	$2P_{11}P_{21}$	$2P_{11}P_{22} + 2P_{12}P_{21}$	$2P_{12}P_{22}$
A_2A_2	P_{21}^2	$2P_{21}P_{22}$	P_{22}^2

In addition we define X_{ij}'s as unambiguous counts of gametes:

$$X_{11} = 2N_{11} + N_{12} + N_{21}$$
$$X_{12} = 2N_{13} + N_{12} + N_{23}$$
$$X_{21} = 2N_{31} + N_{21} + N_{32}$$
$$X_{22} = 2N_{33} + N_{23} + N_{32}$$

Hill (1974) obtained the maximum likelihood estimate for the frequency of the A_1B_1 gamete (P_{11}) by formulating the likelihood function for getting the observed data given P_{11}, and maximizing the likelihood with respect to P_{11}. This yielded

$$\langle P_{11} \rangle = \frac{[X_{11} + N_{22}P_{11}(1 - \langle p \rangle - \langle q \rangle + P_{11})]/[P_{11}(1 - \langle p \rangle - \langle q \rangle + P_{11}) + (\langle p \rangle - P_{11})(\langle q \rangle - P_{11})]}{2N}$$

where

$$\langle p \rangle = (X_{11} + X_{12} + N_{22})/2N$$

and

$$\langle q \rangle = (X_{11} + X_{21} + N_{22})/2N$$

are the allele frequencies of A_1 and B_1 respectively.

To find the maximum likelihood estimate of D, an initial guess for P_{11} is inserted into the above equation, and it yields a revised $\langle P_{11} \rangle$. This is then plugged back into the right hand side of the same equation, and this process is iterated until the value converges. The maximum likelihood estimate of P_{11} is then substituted into

$$\langle D \rangle = \langle P_{11} \rangle - \langle p \rangle \langle q \rangle$$

to yield the maximum likelihood estimator for D. Weir and Cockerham (1979) show that the iterative process does not always have nice convergence properties, and that by directly solving the cubic equation for P_{11}, one can avoid this problem.

With more than two alleles of a gene, there are several ways to represent the statistical associations among alleles in gametes. A reasonable extension of the diallelic case is to define a disequilibrium parameter for each gametic type:

$$D_{ij} = P_{ij} - p_i q_j \tag{1.20}$$

where P_{ij} is the frequency of the gamete with alleles A_i and B_j, and the frequencies of these alleles are p_i and q_j respectively. One particular problem with D_{ij} defined in this way is that it is constrained by the allele frequencies. Measures of linkage disequilibrium that are frequency dependent are less useful than statistics that do not depend on allele frequencies for comparison of measures of D for different allele pairs. With a frequency dependent measure, the observation of one D_{ij} value being larger than another would not necessarily imply that the former pair of alleles was more strongly associated. One means of reducing the frequency dependence is to use the measure

$$D_{ij}' = D_{ij}/D_{\max} \tag{1.21}$$

so that each D_{ij} is normalized to the maximum value it could have. Regardless of allele frequency, D_{ij}' varies between -1 and 1. Properties of several other measures of multiple allele linkage disequilibrium and multiple locus linkage disequilibrium are explored by Bennett and Oertel (1965); Smouse (1974); Weir and Cockerham (1978); and Hedrick (1987).

PROBLEMS

1 Show that a diploid organism with n alleles can have $[n(n + 1)]/2$ genotypes. What is the analogous number in a tetraploid?

2 Levy and Levin (1975) used electrophoresis to study the gene for phosphoglucose isomerase-2 in the evening primrose *Oenothera biennis*, a complex genomic heterozygote made permanent by chromosomal translocations. They observed two alleles affecting electrophoretic mobility of the enzyme, and among 57 strains they found 35 *PGI-2*^{a}/*PGI-2*^{a}, 19 *PGI-2*^{a}/*PGI-2*^{b}, and 3 *PGI-2*^{b}/*PGI-2*^{b} genotypes.
   a. Calculate the allele frequencies of *PGI-2*^{a} and *PGI-2*^{b}.
   b. What numbers of the genotypes would be expected with random mating?

3 Phenylketonuria is an autosomal recessive form of severe mental retardation. About 1 in 10,000 newborn Caucasians are affected. Assuming random mating, what is the frequency of heterozygous carriers?

4 Mourant et al. (1976) cite data on 400 Basques from Spain, of which 230 were Rh^+ and 170 were Rh^-. Estimate the allele frequencies of D and d. How many of the Rh^+ individuals are expected to be heterozygous?

5 Kelus (cited in Mourant et al. 1976) reports a study of 3100 Poles, of whom 1101 were MM, 1496 were MN, and 503 were NN. Calculate the allele frequencies and the expected numbers of the three genotypes and carry out a χ^2 test for goodness-of-fit to random-mating proportions.

6 Consider an autosomal gene with four alleles A_1, A_2, A_3, and A_4 with respective frequencies 0.1, 0.2, 0.3, and 0.4. Calculate the expected genotype frequencies with random mating.

7 Show that the proportion of heterozygous offspring from a heterozygous parent is 1/2 in a population undergoing random mating for a single gene with two alleles.

8 If random mating with two alleles gives the frequencies D, H, and R for homozygous dominant, heterozygote, and homozygous recessive, show that $DR = H^2/4$.

9 When mating is random for a gene with two alleles A and a at frequencies p and q, show that the genotype frequencies of AA, Aa, and aa are approximately $1 - 2q$, $2q$, and 0 when q is so small that $q^2 \approx 0$.

10 In a population undergoing random mating for a single gene with a dominant and recessive allele, show that the allele frequency of the recessive allele among individuals with the dominant phenotype is $q/(1 + q)$, where q is the allele frequency of the recessive in the whole population.

11 The frequency of one form of recessive X-linked color blindness is 5% among European males. What is the expected frequency of this form of color blindness among females? What fraction of females would be heterozygous carriers?

12 In a sample of blood from African Pygmies, the counts of the phenotypes were: O, 88; A, 44; B, 27; and AB, 4 (Cavalli-Sforza and Bodmer 1971).
   a. What are the preliminary estimates of $\langle p \rangle$, $\langle q \rangle$, and $\langle r \rangle$?
   b. What are the estimates of genotype frequencies?
   c. Assuming Hardy-Weinberg proportions, what are the expected numbers of O, A, B, and AB phenotypes?
   d. What are the revised allele frequencies after the M step (Box H, p. 40)?
   e. Use a χ^2 test (there is one degree of freedom; why?) to determine whether the assumption of random mating seems justified.

13 For a trait due to a rare X-linked recessive gene, show that the frequency of carrier females is approximately equal to two times the frequency of affected males.

14 What is the analogue of the Hardy-Weinberg principle for a gene with two alleles in a tetraploid?

15 Given the following table of allele frequencies,

		Gene				
		1	2	3	4	5
Allele	1	0.63	0.94	0.995	1.0	0.78
	2	0.37	0.06	0.005	—	0.12
	3	—	—	—	—	0.06
	4	—	—	—	—	0.04

what is the proportion (P) of polymorphic genes (using the definition in the text)? Assuming random mating and linkage equilibrium, what is the average heterozygosity (H) for the set of genes?

16 In a restriction fragment analysis of the gene for 6-phosphogluconate dehydrogenase in 10 naturally occurring strains of *Escherichia coli*, 14 cleavage sites were found using eight enzymes with 6-base restriction sites, and the number of sequences cut at each site was as follows (data from Hartl et al. unpublished):

Site	1	2	3	4	5	6	7	8	9	10	11	12	13	14
No. cut (c_i)	1	3	2	1	9	3	9	1	6	7	10	1	2	1

Use Equations 1.1 and 1.2 to estimate the proportion of polymorphic nucleotide sites in this gene $\langle P_{nuc} \rangle$ and the heterozygosity of nucleotide sites in the genes taken pairwise $\langle H_{nuc} \rangle$. (The heterozygosity is a sort of "virtual heterozygosity," since *E. coli* is haploid.)

17 Charles Darwin could have discovered segregation had he known what to look for, as Mendelian segregation occurred in at least one of his own experiments. Darwin (cited in Iltis 1932) studied flower shape in the snapdragon *Antirrhinum*. In a cross between a true-breeding strain with regular (peloric) flowers and a true-breeding strain with irregular (normal) flowers, all of the F_1's were normal. Crosses of $F_1 \times F_1$ yielded 88 normal and 37 peloric plants. Perform a χ^2 test assuming a 3:1 ratio in the F_2. Is the peloric or normal allele dominant?

18 For a mating between triple dominant/recessive heterozygotes of three unlinked genes, there are eight phenotypic classes among the offspring. What are the expected phenotypic ratios? Mendel carried out such an experiment and obtained the phenotypic ratio 269:98:86:88:30:34:27:7 among a total of 639 progeny. (He complained that this experiment required the most time and effort of any of his crosses.) Calculate the χ^2 and associated probability.

19 A cross of *Ab/aB* with *ab/ab* produced 60 nonrecombinant and 40 recombinant offspring. Is there evidence of linkage?

20 For the two genes in Problem 19, the cross *AB/ab* with *ab/ab* gave the following progeny: 84 *AB/ab*, 39 *Ab/ab*, 37 *aB/ab*, 80 *ab/ab*.

a. Are these data consistent with a recombination fraction of 0.40?

b. Combining these results with those in 18, what is the best estimate of the recombination fraction between the *A* and *B* genes?

21 If one gene has alleles A_1 and A_2 at frequencies p_1 and p_2, and another gene has alleles B_1, B_2, and B_3 at frequencies q_1, q_2, and q_3, what are the expected frequencies of gametes with linkage equilibrium, assuming that $p_1 = 0.3$, $q_1 = 0.2$, and $q_2 = 0.3$?

22 For two genes with alleles A_1 and A_2 and B_1 and B_2, respectively, with p_1 and p_2 the allele frequencies of A_1 and A_2, and q_1 and q_2 those of B_1 and B_2, let $p_1 = 0.7$ and $q_1 = 0.3$.

a. What are the frequencies of all possible gametes assuming linkage equilibrium?

b. What are the frequencies of all possible gametes, if there is linkage disequilibrium with *D* equal to 50% of its theoretical maximum?

23 Use the result in Problem 10 to show that the frequency of homozygous recessive genotypes from dominant × dominant matings is $[q/(1 + q)]^2$ and from dominant × recessive matings is $q/(1 + q)$. Note that the latter is equal to the square root of the former. (These proportions are called Snyder's ratios and were once used to test traits for simple recessive inheritance.)

24 Show that the number of possible genotypes in a *p*-ploid with *n* alleles is $(p + n - 1)!/[p!(n - 1)!]$.

Chapter Two

Random Genetic Drift

The Hardy-Weinberg principle provides a static framework for the organization of genetic variation in populations. In the absence of any outside forces to change allele frequency, the genotype frequencies settle to the products of the component allele frequencies (with an extra factor of two in heterozygotes). The Hardy-Weinberg principle is very robust, and deviations from it are often considered in theoretical and experimental population genetics. In the formal theory of statistical hypothesis testing, the investigator constructs a **null hypothesis** (typically, that the data fit a model perfectly); a test is then performed to see whether the observed data deviate from the expectations of the null hypothesis only as a result of chance sampling. For example, in testing the fairness of a coin, the null hypothesis is that half the time the coin will turn up heads and half the time tails, although by chance alone a fair coin could yield 55 heads in 100 tosses with reasonable probability. A population with 100 individuals undergoes an analogous random process in each generation when gametes unite in the process of reproduction. Changes in allele frequency can result from chance alone, and because the allele frequencies do not change in any predetermined way by the sampling process, the chance process is known as **random genetic drift**. All other forces that can change gene frequency — mutation, natural selection, and migration — should really be considered in concert with random genetic drift, because the sampling process occurs in every population. As we will see, the magnitude of allele frequency change each generation depends on the population size, and random genetic drift becomes less important in large populations. The effect of population size on processes that change gene frequencies is the issue that Wright and Fisher most disagreed on: Fisher viewed populations as sufficiently large that random genetic drift could be ignored, while Wright, who was experienced with the genetic effects of small populations frequently encountered in agricultural genetics, was more concerned with chance sampling.

BINOMIAL SAMPLING

If you toss a coin 10 times there are 11 possible outcomes, not counting different orders of heads and tails. You could obtain 0, 1, 2, . . . , or 10 heads, with the rest tails. The probability of each of the 11 possibilities is given by the binomial distribution discussed below. In sampling gametes from a finite population, the sampling process is more like drawing different colored marbles from an urn, with two different colors representing two different alleles. We view the urn as containing an essentially infinite number of marbles, just as drift theory assumes that gametes are sampled from a very large number. The critical point is we assume that as gametes or marbles are drawn, the frequencies of those remaining do not change. This assumption is equivalent to a process called "sampling with replacement," and it also results in a binomial distribution of outcomes, just as in coin tossing.

To take a specific example, a population of 9 diploid organisms arises from a sample of just 18 gametes, but the gametes can be thought of as being sampled from an essentially infinite pool. Because small samples are frequently not representative, an allele frequency in the sample may differ from that in the entire pool of gametes. In fact, if the number of gametes in a sample is represented as $2N$ (in this example, $2N = 18$), the probability that the sample contains exactly i alleles of type A is the binomial probability

$$\Pr(i) = \binom{2N}{i} p^i q^{2N-i} \qquad (2.1)$$

where $\binom{2N}{i}$ means $(2N)!/i!(2N - i)!$; p and q are, respectively, the allele frequencies of A and a in the entire pool of gametes ($p + q = 1$); and i takes on any integer value between 0 and $2N$. The new allele frequency in the population (call it p') is therefore $i/2N$ because, by definition, the allele frequency of A equals the number of A alleles (in this case i) divided by the total (in this case $2N$). In the next generation, the random sampling process occurs again, and the new probability of a prescribed number of A alleles occurring in the $2N$ gametes is given by the binomial probability above, with p now replaced by p' and q by $1 - p'$. Thus, the allele frequency may change at random from generation to generation. Computer-generated examples based on random numbers are shown in Figure 1. Each line in Figure 1a gives the number of A alleles in 20 successive generations of random genetic drift in a population of size $N = 9$ (so $2N = 18$). As you can see, individual populations behave very erratically. In 7 populations, the A allele became **fixed** (that is, $p = 1$); in 5 populations, A was **lost** (that is, $p = 0$). The other 8 populations remained **unfixed** (A was neither fixed nor lost), but the final allele frequency among the unfixed populations was as likely to be one value as another. Figure 1b shows the same kind of simulation, except here $2N = 100$. With a larger population size, the rate at which populations go to

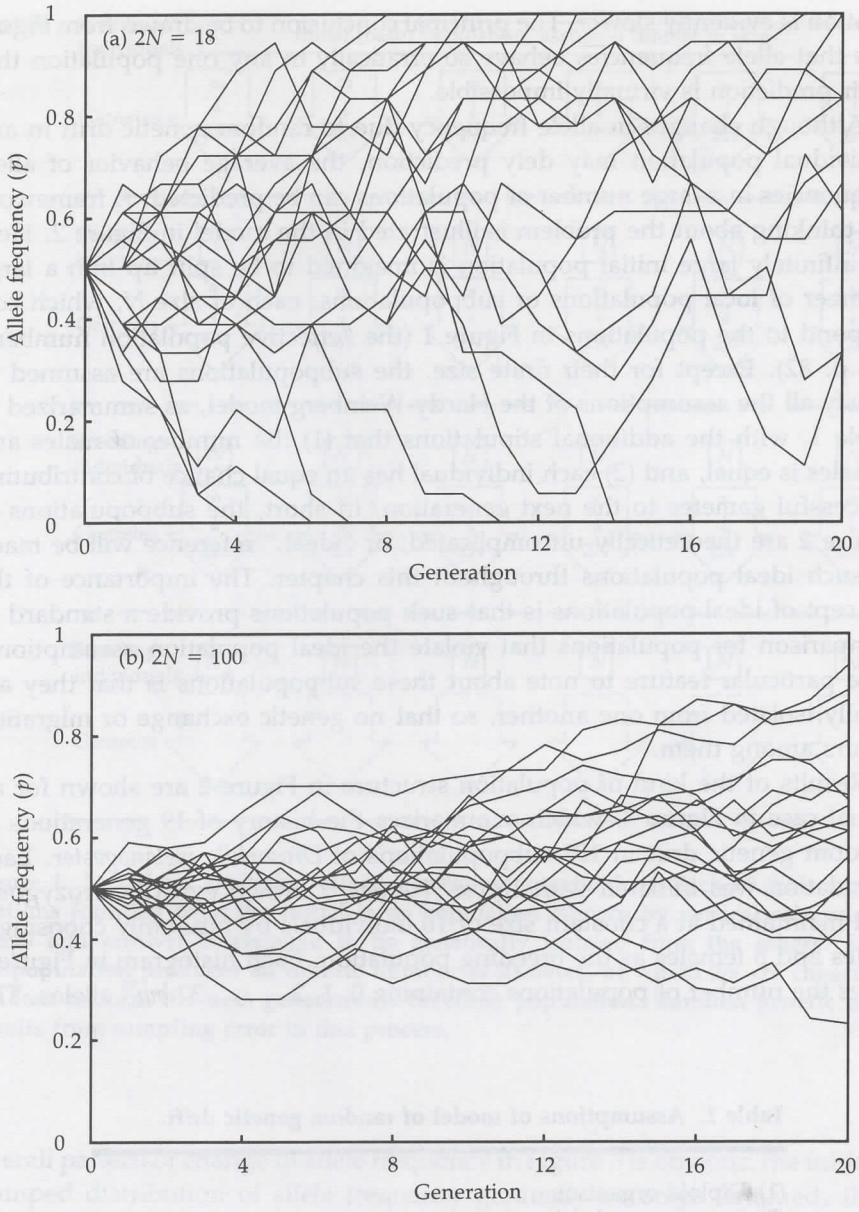


Figure 1. Computer simulations of the Wright-Fisher model of random genetic drift. Each line in (a) and (b) represents a population size of $2N = 18$ or $2N = 100$, respectively, simulated for 20 generations. Each generation alleles are sampled with replacement as described in the text. An allele frequency of $p = 0.5$ in (a) implies that there are 9 copies of the *A* allele and 9 copies of the *a* allele. In (b), an allele frequency of 0.5 implies 50 copies of each allele. Note that the larger population size in (b) results in smaller oscillations of allele frequency, and a slower rate of fixation.

fixation is evidently slower. The principal conclusion to be drawn from Figure 1 is that allele frequencies behave so erratically in any one population that their prediction is virtually impossible.

Although changes in allele frequency due to random genetic drift in any individual population may defy prediction, the average behavior of allele frequencies in a large number of populations can be predicted. A framework for thinking about the problem is illustrated in the model in Figure 2. Here an infinitely large initial population is imagined to be split up into a large number of local populations or subpopulations, each of size N, which correspond to the populations in Figure 1 (the "effective population number;" see p. 82). Except for their finite size, the subpopulations are assumed to satisfy all the assumptions of the Hardy-Weinberg model, as summarized in Table 1, with the additional stipulations that (1) the number of males and females is equal, and (2) each individual has an equal chance of contributing successful gametes to the next generation. In short, the subpopulations in Figure 2 are theoretically uncomplicated, or "ideal;" reference will be made to such ideal populations throughout this chapter. The importance of the concept of ideal populations is that such populations provide a standard of comparison for populations that violate the ideal population assumptions. One particular feature to note about these subpopulations is that they are totally isolated from one another, so that no genetic exchange or migration occurs among them.

Results of the kind of population structure in Figure 2 are shown for an actual case in Figure 3, which summarizes the history of 19 generations of random genetic drift in 107 subpopulations of *Drosophila melanogaster*. Each population was initiated with 16 bw^{75}/bw (bw = brown eyes) heterozygotes and maintained at a constant size of 16 individuals by randomly choosing 8 males and 8 females as the breeding population. Each histogram in Figure 3 gives the number of populations containing 0, 1, 2, . . . , 32 bw^{75} alleles. The

Table 1. Assumptions of model of random genetic drift.

(1) Diploid organism
(2) Sexual reproduction
(3) Nonoverlapping generations
(4) Many independent subpopulations, each of constant size N
(5) Random mating within each subpopulation
(6) No migration between subpopulations
(7) No mutation
(8) No selection

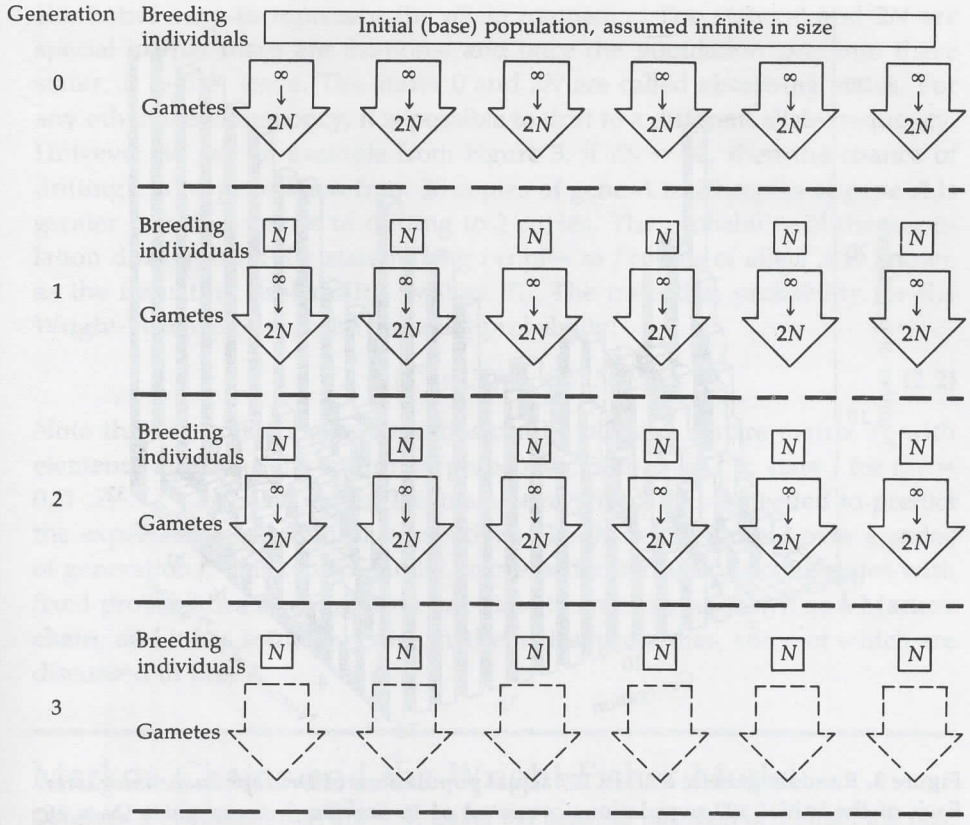


Figure 2. Model for analyzing effects of random genetic drift. Each of the subpopulations founded from the initial large population (shown by vertical columns of boxes and arrows) is assumed to be genetically isolated from the others. Each subpopulation produces an infinite number of gametes, of which 2*N* are chosen at random to form the next generation's breeding population. Random genetic drift results from sampling error in this process.

overall pattern of change in allele frequency in Figure 3 is obvious: the initially humped distribution of allele frequency gradually becomes flattened, then U-shaped as populations fixed for bw^{75} or *bw* begin to pile up at the boundaries. The piling up occurs because once an allele has been fixed or lost, it remains fixed or lost, since mutation is negligible over such a small number of generations in small populations. By about 18 generations, half the populations are fixed for one allele or the other, and among the unfixed populations, the distribution of allele frequencies is essentially flat.

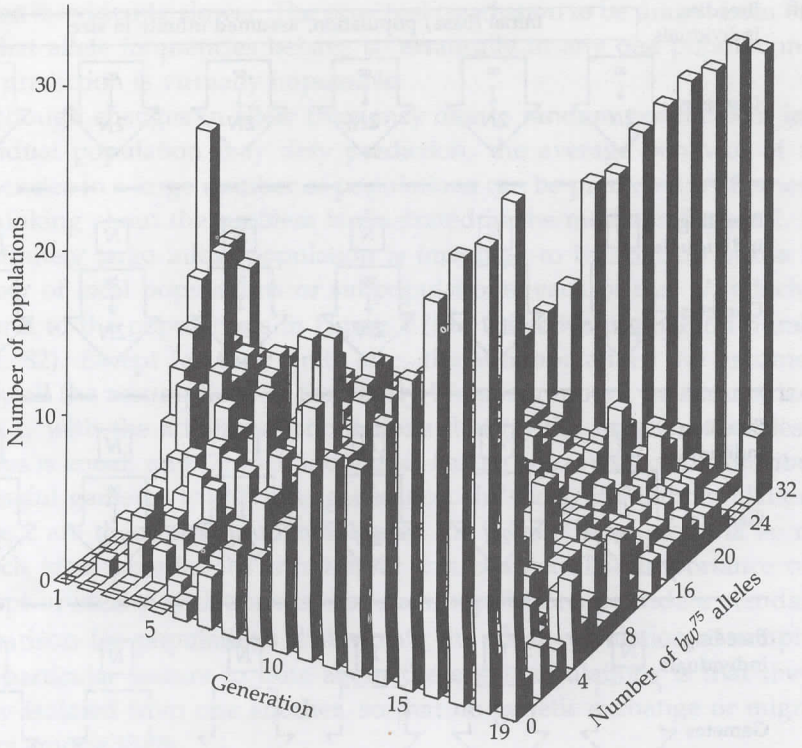


Figure 3. Random genetic drift in 107 actual populations of *Drosophila melanogaster*. Each of the initial 107 populations consisted of 16 bw^{75}/bw heterozygotes (N = 16; bw = brown eyes). From among the progeny in each generation, 8 males and 8 females were chosen at random to be the next generation. As time passes, an increasing number of populations become fixed for one allele or the other. (Data from Buri 1956.)

THE WRIGHT-FISHER MODEL OF RANDOM GENETIC DRIFT

Fisher (1930) and Wright (1931) both considered the consequences of the sort of binomial sampling that occurs in small populations when the sampling occurs repeatedly over many generations. This model, known as either the Wright-Fisher or Fisher-Wright model, derives the distribution of allele frequencies among populations undergoing random genetic drift. Although neither Fisher nor Wright formulated the problem in terms of matrices, as used here (see following and Box A), this approach makes the problem much simpler and gives the same results. If a population has $2N$ genes, and there are two alleles (A and a) that may be segregating, then the **state** of the population can be described by the number of A alleles in the population. The possible states are then 0, 1, 2, . . . , $2N$. The state is really just an

alternative way to represent the allele frequency. The states 0 and $2N$ are special in that these are fixations, and once the population gets into these states, it cannot leave. The states 0 and $2N$ are called **absorbing states**. For any other allele frequency, it is possible to drift to a different allele frequency. However, to use an example from Figure 3, if $2N = 32$, then the chance of drifting in one generation from 30 copies of gene A to 29 copies of gene A is greater than the chance of drifting to 2 copies. The probability of the population drifting from the state having i copies to j copies of allele A is known as the **transition probability**, written T_{ij}. The transition probability for the Wright-Fisher model is the binomial probability:

$$T_{ij} = \binom{2N}{j}(i/2N)^j(1 - i/2N)^{2N-j} \qquad \textbf{(2.2)}$$

Note that the transition probabilities can be put in a square matrix **T**, with elements T_{ij} giving the transition probability from state i to state j for $i, j = 0, 1, 2, \ldots, 2N$. The matrix **T** contains everything that is needed to predict the expected distribution of populations like those in Figure 3 over a series of generations. This type of model, expressed in terms of discrete states with fixed probabilities of going from one state to another, is known as a **Markov chain**, and it has some very nice mathematical properties, some of which are discussed in Box A.

Markov Chains and the Wright-Fisher Model

A

Consider a population with $2N = 4$. Using the binomial definition of T_{ij} from the text, we get the following of the transition probability matrix:

		Number of *A* alleles in generation $t + 1$				
		0	**1**	**2**	**3**	**4**
Number of *A* alleles in generation *t*	**0**	1	0	0	0	0
	1	0.316	0.422	0.211	0.047	0.004
	2	0.062	0.25	0.375	0.25	0.062
	3	0.004	0.047	0.211	0.422	0.316
	4	0	0	0	0	1

Note that the sum of the elements of each row of the matrix is 1. This occurs because, if the population is in state i in generation t, then at generation $t + 1$ it must be in either state 0, 1, 2, 3, or 4. The probabilities of all possible transitions must sum to 1. Each row T_{i0} to T_{i4} gives the probability distribution of populations starting from state i after a single generation of drift. States 0 and 4 are unusual in that the only nonzero transition probability has a value of 1. This means that from state 0, the only state the population can go to is 0 ($T_{00} = 1$). In the same manner, the population can be "absorbed" by state 4. States 0 and 4 are the fixations.

The frequency distribution of populations having specified allele frequencies can be obtained very quickly from the Markov chain formulation of the Wright-Fisher model.

Define X_i as the frequency of populations having i copies of the A allele, or more precisely, being in state i (i = 0, 1, 2, 3, or 4), and let **X** be the vector of X_i's. The frequency distribution after one generation of random drift is determined by matrix multiplication:

$$\mathbf{X}' = \mathbf{XT}$$

where, by the usual definition of matrix multiplication,

$$X_j' = \Sigma_i X_i T_{ij}$$

The calculations are tedious, but a computer can be programmed to do the task. Markov chains afford easy computation of the steady-state or stationary distribution, and the eigenvalues of the matrix tell us about the rate at which the populations change. The largest eigenvalue of the transition probability matrix for the Wright-Fisher model is

$$\lambda = 1 - 1/2N$$

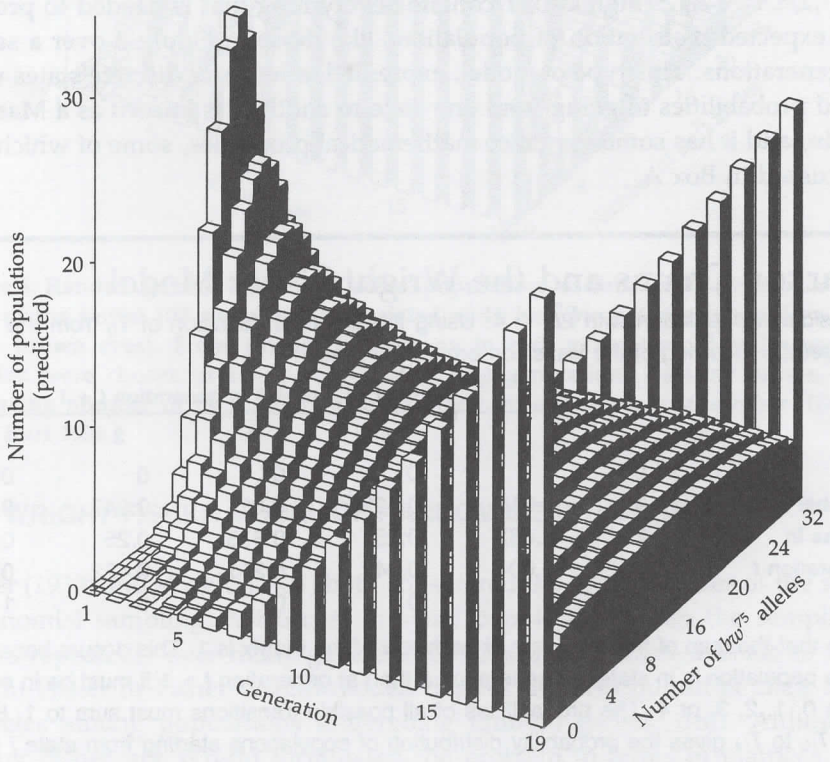


Figure 4. Prediction of the Wright-Fisher model for the distribution $\phi(x,t)$ of populations of size N = 16, with allele frequency x at generation t, for 19 generations after an initial frequency of 0.5. The values of $\phi(x,t)$ were generated using the Markov transition probability matrix. The model with $2N$ = 32 predicts that fewer populations have fixed by generation 19 than actually did go to fixation in the experiment. This is because the effective population size is smaller than the observed count (see Figure 10).

and this is approximately the rate at which the average heterozygosity among the subpopulations decreases:

$$H_t = \lambda^t H_0$$

For larger populations, a larger number of elements are in the transition probability matrix, and here is where the limitation of the method lies. The transition matrix grows as the square of the population size, and even populations of just 50 individuals require a matrix with 10,201 entries. For the case of $2N = 32$, the calculations can still be performed on a microcomputer, and the iterations of the transition matrix for $2N = 32$ starting with $X_{16} = 1$ (all populations at initial allele frequency 0.5) are plotted in Figure 4. This is the exact theoretical prediction for the experiment done by Buri (1956); see also Figure 3 in this chapter.

One important result from Markov chain analysis is the probability of fixation of an allele. Markov chain theory provides the probability of being in a particular state at a particular time. Under certain conditions Markov chains converge to a stationary distribution in which the probability of being in state i at generation t remains constant for all t. As in Box A, the probability distribution can be obtained by matrix multiplication, and the stationary probability distribution must satisfy

$$\pi_j = \Sigma_i \pi_i T_{ij} \qquad \textbf{(2.3)}$$

Suppose that the initial frequency of allele A is π_0. The unique stationary probability distribution for the Markov chain that specifies the Wright-Fisher model is

$$\boldsymbol{\pi} = (1 - \pi_0, 0, 0, 0, \ldots, \pi_0) \qquad \textbf{(2.4)}$$

where the number of 0's equals $2N - 1$. This stationary distribution says that, after a sufficiently long time, the probability of being in state 0 (0 copies of A) is $1 - \pi_0$, and the probability of being in state $2N$ (fixation of allele A) is π_0. In other words, *the probability of ultimate fixation is simply the initial frequency of the allele.* Another means of obtaining this result was suggested by Ewens (1979) and does not require any heavy mathematics. It rests on the notion of the coalescent, which will be discussed in more detail in the last section of this chapter. The basic idea is that after a long enough time has passed, every gene in the population must have descended from just one gene in the initial population. All other allele lineages will have died out. The probability that the sole surviving lineage is from gene A is equal to the initial frequency of gene A; hence this is the fixation probability of A.

Fisher and Wright extended their model to include the effects of mutation and selection. The rationale behind these extensions was to learn how the forces of mutation and selection would affect the transition probability matrix,

and to modify the matrix accordingly. For example, under bidirectional mutation, the two fixation states would no longer be absorbing. Selection also affects the elements of the transition probability matrix, but they are still constants, and the standard theory of Markov chains still applies. Fisher and Wright also addressed the expected time to fixation. Since the diffusion approximation yields the solution to this problem much more easily, we will consider times to fixation in the next section.

So far genetic drift has been considered as occurring due to the sampling of gametes from an essentially infinite pool. In some organisms the sampling really occurs in the zygotic or genotypic stage. For example, some shellfish produce vast numbers of pelagic larvae that drift about in the sea, with only a small fraction surviving to adulthood. This type of sampling might seem to be different from gametic sampling, since there are three genotypes to be drawn from a large pool. If the frequencies of the three genotypes are x, y, and z, where $x + y + z = 1$, and the successive samples are independent of one another, then the distribution of the sampling process is trinomial, written as

$$\Pr(n_1, n_2, n_3) = \frac{N!}{n_1!n_2!n_3!} x^{n_1} y^{n_2} z^{n_3} \tag{2.5}$$

This distribution gives the probability of obtaining a sample of n_1, n_2, and n_3 of genotypes AA, Aa, and aa, respectively, where $n_1 + n_2 + n_3 = 1$. Note that the trinomial form is very similar to the binomial. If there were m genotypes, the multinomial distribution would be

$$\Pr(n_1, n_2, \ldots n_m) = \frac{N!}{n_1!n_2! \ldots n_m!} x_1^{n_1} x_2^{n_2} \ldots x_m^{n_m} \tag{2.6}$$

This multinomial distribution gives the probability of drawing a sample composed of n_i individuals of the genotypes with frequency x_i, where $i = 1, 2, \ldots, m$. At this point the Hardy-Weinberg principle becomes important in the theory of random drift. The sampling of genotypes is really a sort of double sampling of gametes because each genotype represents two gametes. In fact, if the pairing of gametes into zygotes is strictly independent, then genotype sampling is formally the same as gametic sampling. The Hardy-Weinberg principle assumes that genotypes are formed by independent sampling of gametes, so under the condition that the population is in Hardy-Weinberg equilibrium, sampling of gametes has the same drift properties as the sampling of genotypes. Departures from Hardy-Weinberg proportions require that multinomial sampling be considered, which can result in a covariance between gametes, but the details need not concern us here.

THE DIFFUSION APPROXIMATION

The pattern of change in allele frequency shown in Figure 3 is very nearly that expected theoretically for an ideal population. Although the full-blown theory of random genetic drift requires mathematics beyond the scope of this book, some background might be of interest (see Kimura 1955, 1964, 1976; Crow and Kimura 1970; Kimura and Ohta 1971; Wright 1969). The representation of random drift by a differential equation was first applied by Fisher (1922), who noted that the equation describing the diffusion of heat through a solid bar applies to random genetic drift. The distribution of populations with allele frequencies ranging from 0 to 1 is called $\phi(x,t)$, where x now represents the allele frequency and t indicates time. Figure 3 shows an experimental realization of $\phi(x,t)$ changing through time. The theoretical problem is to formulate an equation that describes how $\phi(x,t)$ changes under random genetic drift, and to solve the equation. Fisher used the equation for heat diffusion, which has the form

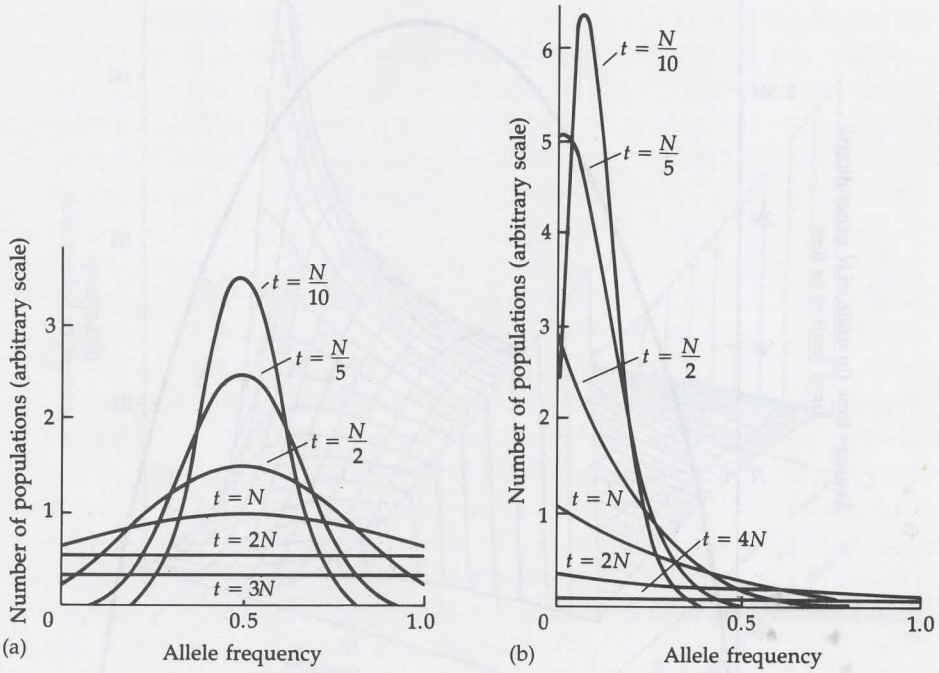


Figure 5. Theoretical results of random genetic drift. (a) Initial allele frequency = 0.5. (b) Initial allele frequency = 0.1. The curves have been scaled so that the area under each curve is equal to the proportion of populations in which fixation or loss has not yet occurred. The curves are therefore the distributions of allele frequencies among segregating populations. (From Kimura 1955.)

$$\frac{\partial \phi(x,t)}{\partial t} = \frac{1}{4N} \frac{\partial^2}{\partial x^2} [x(1 - x)\phi(x,t)] \qquad (2.7)$$

Many aspects of this problem were explored by Wright (1931), and the formal solution to this equation, found by Kimura (1955), requires complex mathematics including Gegenbauer polynomials, the hypergeometric function, and other advanced techniques. For our purposes, a graph of the solution will suffice. The two families of curves in Figure 5 are the theoretical distributions $\phi(x,t)$ of allele frequency among unfixed populations after various times (t) measured in units of N generations. In Figure 5a, all populations have an initial allele frequency of 0.5, as in the actual populations in Figure 3; after about $t = 2N$ generations, the distribution of allele frequency is essentially flat, and by this time about half the populations are still unfixed. The distributions in Figure 5 refer only to those populations that are unfixed; as time

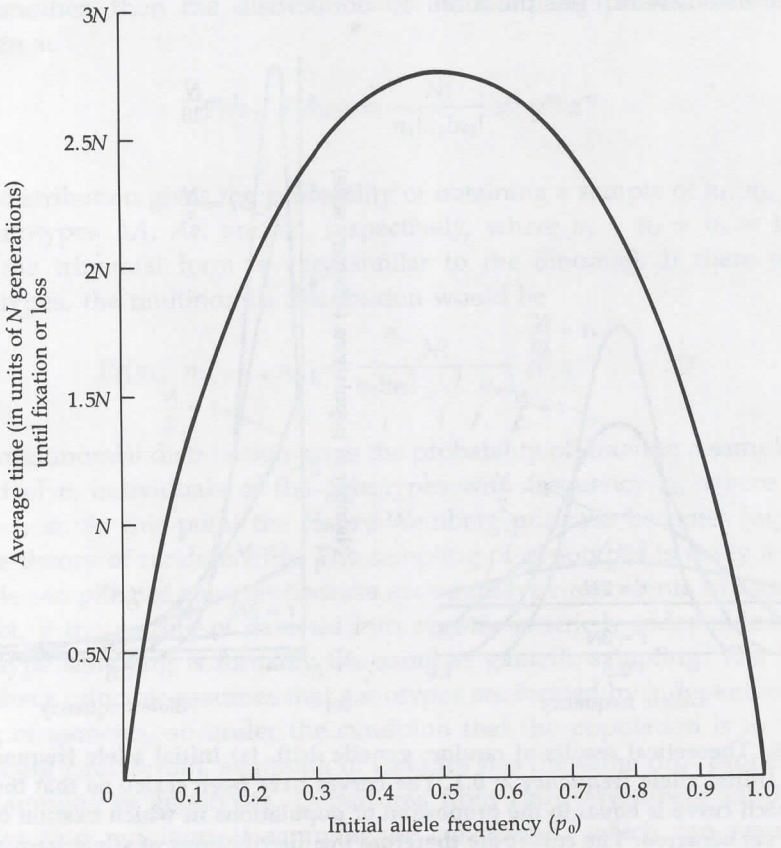


Figure 6. Average persistence of a neutral allele in an ideal population of size N, plotted against initial allele frequency.

goes on, more and more of the populations do become fixed, and the distributions progressively pile up at 0 and 1, as in the histograms in Figure 3. Indeed, in Figure 5, the area under each curve is equal to the proportion of unfixed populations, which becomes progressively smaller. In particular, the rate at which the height of the distribution decreases once it becomes flat is about $1/2N$ per generation.

Figure 5b shows what happens when the initial allele frequency is 0.1; here the distributions are highly asymmetrical, and the distribution of allele frequency does not become flat until about $t = 4N$ generations, by which time only about 10% of the populations remain unfixed. Once a flat distribution of allele frequency is reached, the distribution remains flat, but random drift continues until fixation or loss has occurred in all populations. The average time required for fixation or loss is shown in Figure 6, which gives the average time in units of N generations that an allele, initially at frequency p_0, remains in the population before it is fixed or lost. When $p_0 = 0.5$, the

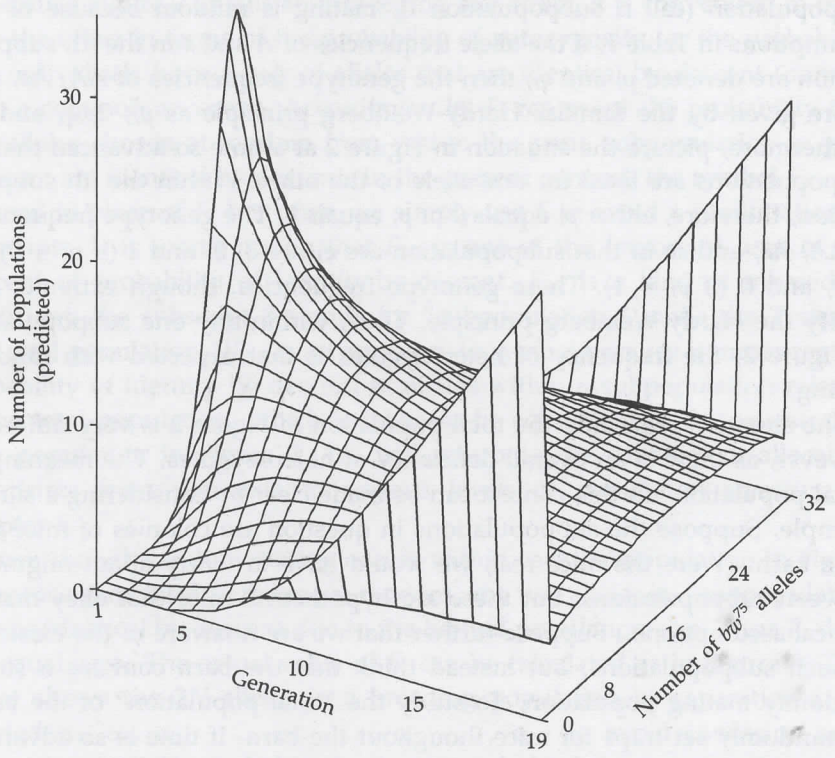


Figure 7. Kimura's (1955) solution to the diffusion equation for the particular case of $N = 16$. This is the three-dimensional view of Figure 5, and represents the diffusion approximation to the exact solution obtained by the Wright-Fisher model in Figure 4.

average time that a population remains unfixed is about $2.8N$ generations. To show that the diffusion approximation and the Wright-Fisher model give very similar results, Figure 7 shows the diffusion approximation for the data in Figure 3, where $2N = 32$, $x_0 = 0.5$, and the value of t is from 1 to 19.

PARALLELS BETWEEN RANDOM DRIFT AND INBREEDING

Many of the important consequences of random genetic drift follow from the fact that the population structure in Figure 2 entails a peculiar sort of "inbreeding," but the nature of the inbreeding is very subtle because the population structure in Figure 2 can be considered on two distinct levels — the level of the individual subpopulation and the level of the total population, which is the aggregate of all subpopulations.

Consider first any one of the subpopulations in Figure 2. Within this subpopulation (call it subpopulation i), mating is random because of the assumptions in Table 1; if the allele frequencies of A and a in the ith subpopulation are denoted p_i and q_i, then the genotype frequencies of AA, Aa, and aa are given by the familiar Hardy-Weinberg principle as p_i^2, $2p_iq_i$ and q_i^2. Furthermore, picture the situation in Figure 2 at a time so advanced that all subpopulations are fixed for one allele or the other. Within the ith subpopulation, therefore, either p_i equals 0 or p_i equals 1. The genotype frequencies of AA, Aa, and aa in that subpopulation are either 0, 0, and 1 (if $p_i = 0$), or 1, 0, and 0 (if $p_i = 1$). These genotype frequencies, though extreme, still satisfy the Hardy-Weinberg principle. Thus, within any one subpopulation in Figure 2, the frequency of heterozygotes is that expected with random mating.

The situation regarding the total population in Figure 2 is very different, however, as there is an overall deficiency of heterozygotes. The meaning of "total population" in this context can be made clear by considering a simple example. Suppose the subpopulations in question are colonies of mice in a large barn. (Were the mice real, we would have to worry about migration between subpopulations, but these are hypothetical mice that obey mathematical assumptions.) Suppose further that we are unaware of the existence of such subpopulations, but instead think that the barn contains a single randomly mating population. To study the "total population" of the barn, we randomly set traps for mice thoughout the barn. If time is so advanced that a fraction p of the subpopulations are fixed for A and a fraction q are fixed for a (remember, $q = 1 - p$), then a fraction p of the time, we will trap an AA mouse and a fraction q of the time, an aa mouse. The overall allele

frequency of A among the trapped animals would then be p, and we would naively expect a fraction $2pq$ of the animals to be heterozygous. In fact, we would have caught no heterozygotes at all!

This rather paradoxical result — that there is a deficiency of heterozygotes in the total population even though random mating occurs within each subpopulation — is a consequence of the random genetic drift of allele frequencies among subpopulations due to their finite size. Were the subpopulations so large that random drift could be ignored, each subpopulation would have the same allele frequencies and Hardy-Weinberg genotype frequencies as any other. In such a case, to return to the mouse example, it would not matter to the proportion of heterozygotes if mice were sampled from the total population or from any one of the subpopulations. For the model in Figure 2, however, random genetic drift is important. The total population has a deficiency of heterozygotes, much as if there were inbreeding. It is this inbreeding-like effect of population subdivision that we now set out to quantify.

Because Figure 2 entails an inbreeding-like effect, it is convenient to measure the effect in terms of the probability of **autozygosity**, or the probability that individuals have a pair of alleles that are identical by descent (derived from a common ancestor). Accordingly, let F_t represent the probability that two alleles chosen at random from within the same subpopulation in generation t are identical by descent. In the present context, the symbol F_{ST} can be used in place of F, but often we simply use F to avoid a proliferation of subscripts. It is worth noting that F_{ST} is one of the important uses of the concept of probability of identity by descent. F_{ST} is a kind of inbreeding coefficient: the subscript S stands for "subpopulation," while the T stands for "total population." F_{ST} is an appropriate symbol because it measures the probability of identity by descent of alleles within a subpopulation relative to the total population, which is taken to be equivalent to the ancestral or base population in Figure 2. We will return to the problem of allocating departures from panmixia to various levels of population structure in Chapter 6.

Even though random mating occurs within each subpopulation in Figure 2, because gametes do combine at random, any two alleles in a subpopulation may be identical by descent due to the limited population size. Thus F_t does not equal zero. The actual value of F_t can be calculated as in Figure 8. This figure shows the $2N$ alleles in a breeding population of generation $t - 1$ (labeled $\alpha_1, \alpha_2, \alpha_3, \ldots, \alpha_i, \ldots, \alpha_j, \ldots, \alpha_{2N}$ to avoid confusion with conventional allele symbols), the gametes derived from generation $t - 1$ (each gametic type representing a fraction $1/2N$ of the entire gametic pool), and two randomly chosen pairs of alleles in the breeding population of

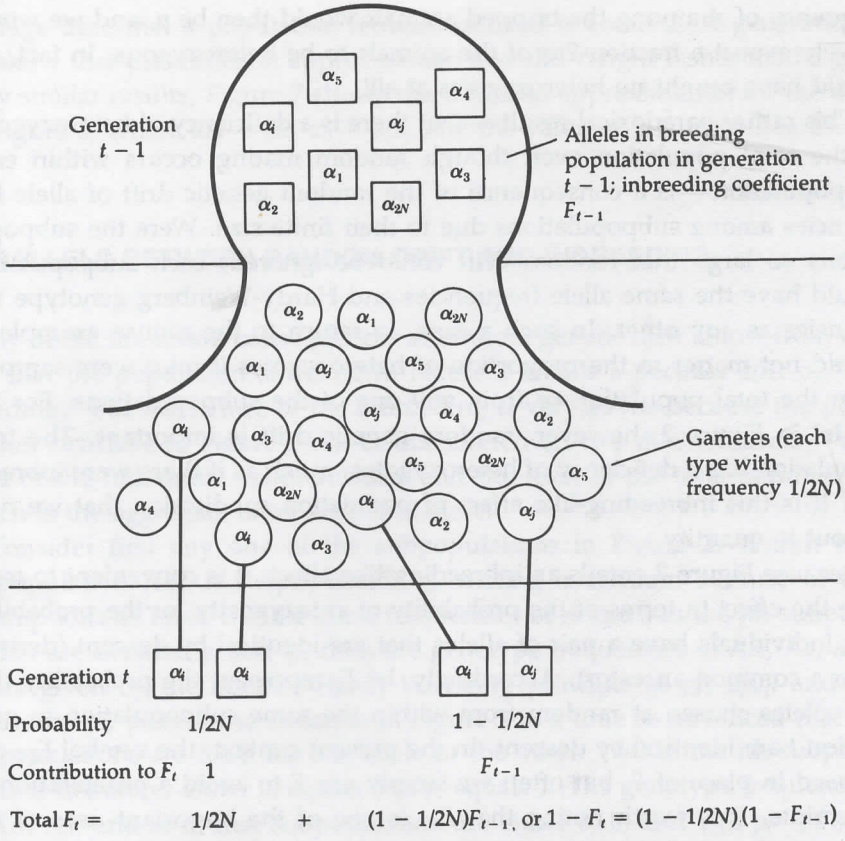


Figure 8. A schematic diagram showing the derivation of the change in autozygosity under random sampling. A pair of gametes can be identical by descent either by being sampled twice (with probability 1/2*N*) or by virtue of previous autozygosity in the population. The recursion for autozygosity is $F_t = (1/2N) + [1 - (1/2N)]F_{t-1}$.

generation t. The probability that the second chosen allele is of the same type as the first is $1/2N$, because this is the frequency of each allelic type in the gametic pool; the probability that the second chosen allele is of a different type from the first is accordingly $1 - 1/2N$. In the first case ($\alpha_i\alpha_i$), the probability of identity by descent is 1; in the second case ($\alpha_i\alpha_j$), it is F_{t-1}. Altogether

$$F_t = 1/2N + (1 - 1/2N)F_{t-1} \tag{2.8}$$

Multiplying both sides by -1 and adding 1 leads to

$$1 - F_t = 1 - 1/2N - (1 - 1/2N)F_{t-1} = (1 - 1/2N)\,(1 - F_{t-1})$$

so

$$1 - F_t = (1 - 1/2N)^t(1 - F_0) \tag{2.9}$$

or simply

$$F_t = 1 - (1 - 1/2N)^t \tag{2.10}$$

when $F_0 = 0$.

Figure 9 shows the rapid increase of F_t in small populations. Another way to see the same phenomenon is to look at the probability of drawing a pair of alleles that are not identical by descent. This probability is the same as the heterozygosity, and can be written

$$H_t = 1 - F_t \tag{2.11}$$

By substituting for F_t we obtain the rate of change in heterozygosity from random genetic drift:

$$H_t = (1 - 1/2N)H_{t-1} \tag{2.12}$$

so

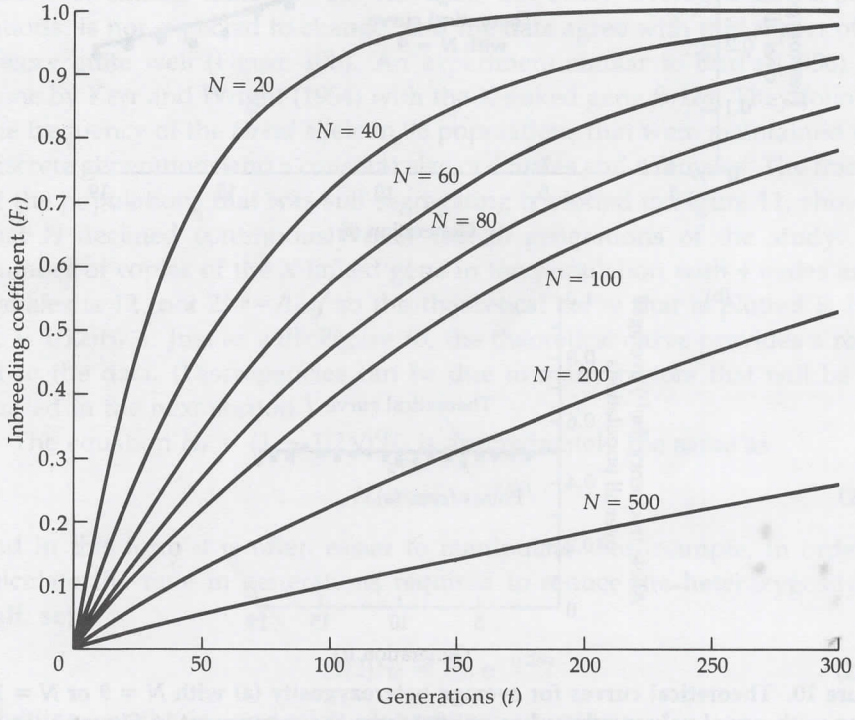


Figure 9. Increase of F_t in ideal populations as a function of time and effective population size N.

$$H_t = (1 - 1/2N)^t H_0 \qquad (2.13)$$

Recall again that a population undergoing random drift remains in approximate Hardy-Weinberg proportions, and that H_t represents a virtual heterozygosity averaged across many subpopulations. The above equations show that pure random drift should result in the heterozygosity decreasing at a geometric rate, since it is multiplied by the constant $(1 - 1/2N)$ each generation. Experimental tests of this prediction are shown in Figures 10 and 11. Figure 10a shows how the heterozygosity averaged across the populations of Buri (1956) declines over generations, but the theoretical curve when $N = 16$ does not fit the data very well. In fact, the rate of decline of hetero-

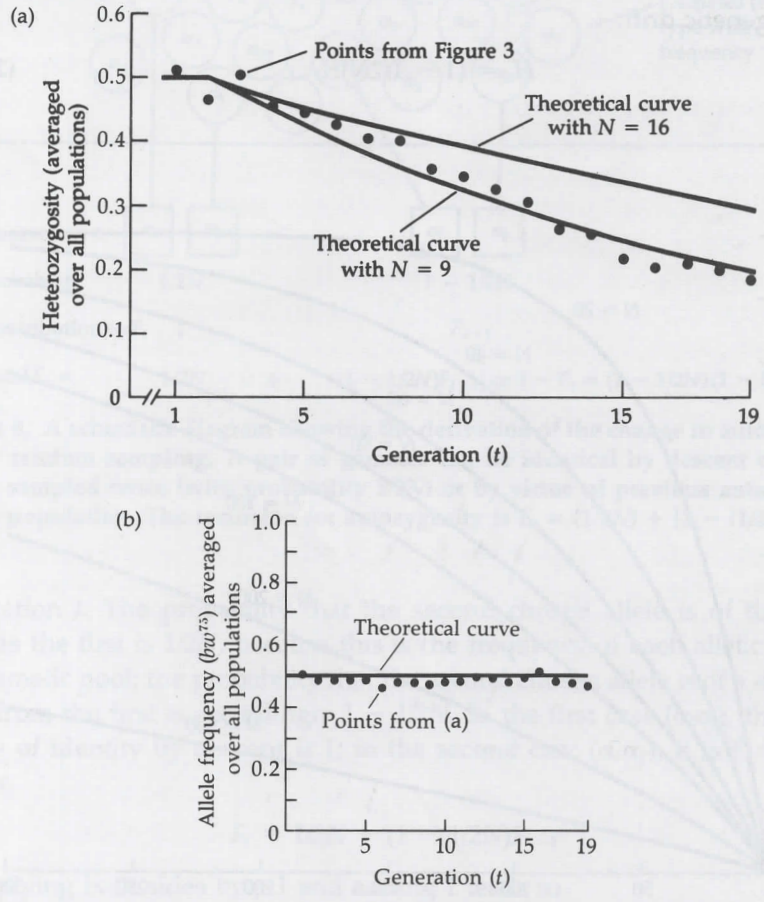


Figure 10. Theoretical curves for average heterozygosity (a) with $N = 9$ or $N = 16$, along with actual values (plotted as points) from the experiment in Figure 3. In (b) the observed and expected allele frequencies (averaged across the 107 populations) are plotted. (Data from Buri 1956.)

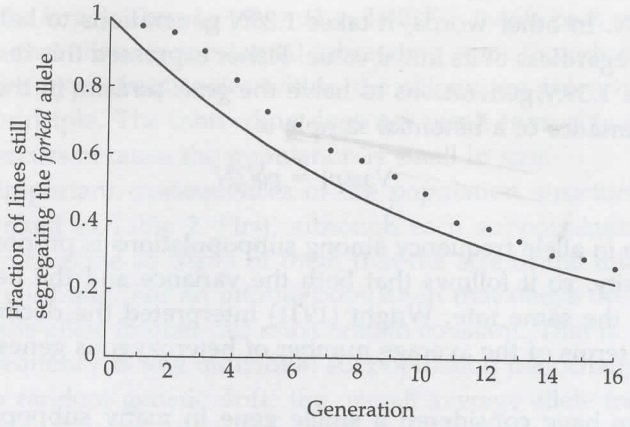


Figure 11. Observed (points) and expected (lines) heterozygosity for an X-linked gene *forked* in an experiment similar to that in Figures 3 and 10. Four males and four females were transferred each generation, and the theoretical recursion is $H_t = (1 - 1/12)H_{t-1}$. (Data from Kerr and Wright 1954.)

zygosity is greater than the theoretical expectation, as though the population size were smaller than $N = 16$. The allele frequency, averaged across populations, is not expected to change, and the data agree with this aspect of the theory quite well (Figure 10b). An experiment similar to Buri's (1956) was done by Kerr and Wright (1954) with the X-linked gene *forked*. They followed the frequency of the *forked* allele in 96 populations that were maintained with discrete generations and a constant size of 4 males and 4 females. The fraction of the populations that was still segregating is plotted in Figure 11, showing that H declined continuously over the 20 generations of the study. The number of copies of the X-linked gene in the population with 4 males and 4 females is 12 (not $2N = 16$), so the theoretical curve that is plotted is $H_t = (1 - 1/12)H_{t-1}$. Just as with Figure 10, the theoretical curve provides a rough fit to the data. (Discrepancies can be due to many factors that will be discussed in the next section.)

The equation $H_t = (1 - 1/2N)^t H_0$ is approximately the same as

$$H_t = H_0 \, e^{-(t/2N)} \qquad \textbf{(2.14)}$$

and in this form it is often easier to manipulate. For example, in order to calculate the time in generations required to reduce the heterozygosity by half, set

$$(1/2)H_0 = H_0 \, e^{-(t/2N)} \qquad \textbf{(2.15)}$$

Dividing out the H_0, and taking logarithms, we obtain

$$\ln(1/2) = -t/2N \qquad \textbf{(2.16)}$$

so $t = 1.39N$. In other words, it takes $1.39N$ generations to halve the heterozygosity, regardless of its initial value. Fisher expressed this result by saying that it takes $1.39N$ generations to halve the *genic variance* in the population. Since the variance of a binomial sample is

$$\text{Var}(p) = pq/2N \tag{2.17}$$

the variance in allele frequency among subpopulations is proportional to the heterozygosity, so it follows that both the variance and the heterozygosity decrease at the same rate. Wright (1931) interpreted the decline in heterozygosity in terms of the average number of heterozygous genes per individual.

So far we have considered a single gene in many subpopulations. We could equivalently consider many genes within a single population. If t is fairly large and the initial allele frequencies of all genes are close to 0.5, then the expression

$$L_t = L_0 \, e^{-(t/2N)} \tag{2.18}$$

gives the approximate number of unfixed genes L_t in the population in generation t.

Strictly speaking, Equation 2.9 applies only to populations that can undergo self-fertilization (many plants, for example), because the theoretical considerations in Figure 8 make no stipulation that half the successful gametes in any generation must come from males and the other half from females. The correction for organisms that cannot undergo self-fertilization is minor, however: if there are equal numbers of males and females, simply replace N in the above equation by $N + 1/2$ (see Crow and Kimura 1970). Since $1/2(N + 1/2)$ is very nearly equal to $1/2N$ for realistic values of N, Equation 2.9 (without the correction) may be taken as an excellent approximation.

As an example of how Equation 2.9 can be used, consider the following question: What is the inbreeding coefficient (i.e., the probability of autozygosity) of a mouse in a colony that was established from a much larger colony 30 generations ago and since that time has maintained a constant size of 20 individuals each generation? For the sake of simplicity, assume that the colony is "ideal" in the sense discussed earlier. The inbreeding coefficient of a mouse is the probability that the alleles of a gene in the mouse are identical by descent; because mating within the colony is random, this probability is equal to the probability that two randomly chosen gametes carry alleles that are identical by descent. The latter probability is, by definition, F_t, so we may use Equation 2.9 with $N = 20$ (as noted above, use of 20.5 would be a bit more accurate), $t = 30$, and $F_0 = 0$ (justified because the base population

was said to be large). Thus $1 - F_t = (1 - 1/40)^{30} = 0.468$, or $F_t = 1 - 0.468 = 0.532$. Note that there is substantial inbreeding even though mating is random and genotypic frequencies within the colony are given by the Hardy-Weinberg principle. The inbreeding does not result from actual nonrandom mating; it occurs because the population is small in size.

Several important consequences of the population structure in Figure 2 are summarized in Table 2. First, although each subpopulation is of finite size, we can imagine so many of them that the size of the total population is effectively infinite. For an infinite population that obeys the assumptions in Table 1, the allele frequencies must remain constant. That is, even though the allele frequency in any individual subpopulation may change any which way due to random genetic drift, the overall average allele frequency of A among subpopulations remains p_0, where p_0 represents the allele frequency of A in the base population. Figure 10b gives an experimental demonstration of the constancy of average allele frequency.

Since F_t is the probability of autozygosity of a gene in an individual in generation t, the probability of allozygosity (obtaining a pair of alleles that are not identical by descent) is $1 - F_t$. Because p_0 is the overall allele frequency of A, the probability that a randomly chosen individual will be genotypically AA is $p_0^2(1 - F_t)$ [for the case of allozygosity] + p_0F_t [for the case of autozygosity]. Similarly, the probability that the individual will be Aa equals $2p_0q_0(1 - F_t)$; and the probability that the individual will be aa equals $q_0^2(1 - F_t) + q_0F_t$. Note that the genotypic frequencies in the total population are different from the standard Hardy-Weinberg proportions, because there is an apparent excess of homozygotes. However, within any one subpopulation, the genotypic frequencies still obey the Hardy-Weinberg principle

Table 2. Consequences of population subdivision and random genetic drift on genotype frequency.

		GENERATION		
		0	t	∞
Inbreeding coefficient (F_t) (Average over all populations)		0	$1 - (1 - 1/2N)^t$	1
Genotype frequency (Average over all populations)	AA:	p_0^2	$p_0^2(1 - F_t) + p_0F_t$	p_0
	Aa:	$2p_0q_0$	$2p_0q_0(1 - F_t)$	0
	aa:	q_0^2	$q_0^2(1 - F_t) + q_0F_t$	q_0
Allele frequency (Average over all populations)	A:	p_0	p_0	p_0
	a:	q_0	q_0	q_0

because of random mating. Substituting for F_t in Equation 2.10 implies that the average heterozygosity among subpopulations at time t equals $2p_0q_0(1 - F_t) = 2p_0q_0(1 - 1/2N)^t$; this is the theoretical curve plotted in Figure 10a and Figure 11 (with $p_0 = q_0 = 0.5$).

Since F_t eventually goes to 1, all subpopulations eventually become fixed for one allele or the other. Because the average allele frequency of A remains p_0 even when all subpopulations have become fixed, the proportion of subpopulations that eventually become fixed for A must be p_0 (and the proportion that eventually become fixed for a must be q_0). Stated another way, the probability of ultimate fixation of an allele in any ideal subpopulation is equal to the frequency of that allele in the initial population. This point is illustrated for an actual example in Figure 3, where $p_0 = 0.5$; by generation 19, 58 populations have become fixed, 30 for the *bw* allele and 28 for bw^{75}.

EFFECTIVE POPULATION SIZE

The effective size of a population usually differs from the actual number of individuals (the effective size is almost always smaller) because no real population obeys all the assumptions in Table 1 exactly. In any actual case, there must be corrections for such complications as fluctuations in population size, unequal numbers of males and females, age structure (Felsenstein 1971; Hill 1972), and unequal family size (for good reviews see Crow and Kimura 1970; Ewens 1982). The degree to which genetic drift can change allele frequencies, and the rates of allele fixation by drift, can be approximated under these complicating circumstances by calculating the effective size of the population and using this value in the theory for an ideal population. That is, the effective population size of an actual population is the number of individuals in a theoretically ideal population having the same magnitude of random genetic drift as the actual population. There are three kinds of effective population size based on either (1) the change in average inbreeding coefficient, (2) the change in variance in allele frequency, or (3) the rate of loss of heterozygosity. These are called the *inbreeding effective size*, the *variance effective size*, and the *eigenvalue effective size*, respectively.

Wright (1931) first worked out the effective population size by considering the effective degree of inbreeding in various situations. As noted, the effective population size can also be calculated by determining the rate of change in variance in a population, and Kimura and Crow (1963) first applied this approach to the problem of overlapping generations. Usually the inbreeding effective size and the variance effective size are the same, but exceptions have been cited (Crow and Kimura 1970). Similarly, the variance effective size and the eigenvalue effective size can be distinct (Ewens 1982). Some of

the factors that require calculation of an effective population size will now be illustrated. We will focus on the inbreeding effective size because this concept is the most widely used.

Fluctuation in Population Size

This is an important example because natural populations actually do fluctuate in size, sometimes by a factor of 10 or more in a single generation (see Figure 12 for two spectacular examples). For the sake of simplicity, assume that the population is ideal in all respects except that its size is not constant. We will consider the situation over just two generations, reserving the more general theory for Box B. Accordingly, suppose that the population's size in two successive generations is N_1 and N_2. The arguments presented in Figure 8 imply that:

$$1 - F_2 = (1 - 1/2N_2)(1 - F_1) \qquad \textbf{(2.19)}$$

and

$$1 - F_1 = (1 - 1/2N_1)(1 - F_0) \qquad \textbf{(2.20)}$$

Substituting from the second equation into the first leads to

$$1 - F_2 = (1 - 1/2N_2)(1 - 1/2N_1)(1 - F_0) \qquad \textbf{(2.21)}$$

By analogy with the constant N case, it is appropriate to try to express this equation in the general form

$$1 - F_t = (1 - 1/2N)^t(1 - F_0) \qquad \textbf{(2.22)}$$

where N is now the effective population size. In our example $t = 2$, so

$$1 - F_2 = (1 - 1/2N)^2(1 - F_0) \qquad \textbf{(2.23)}$$

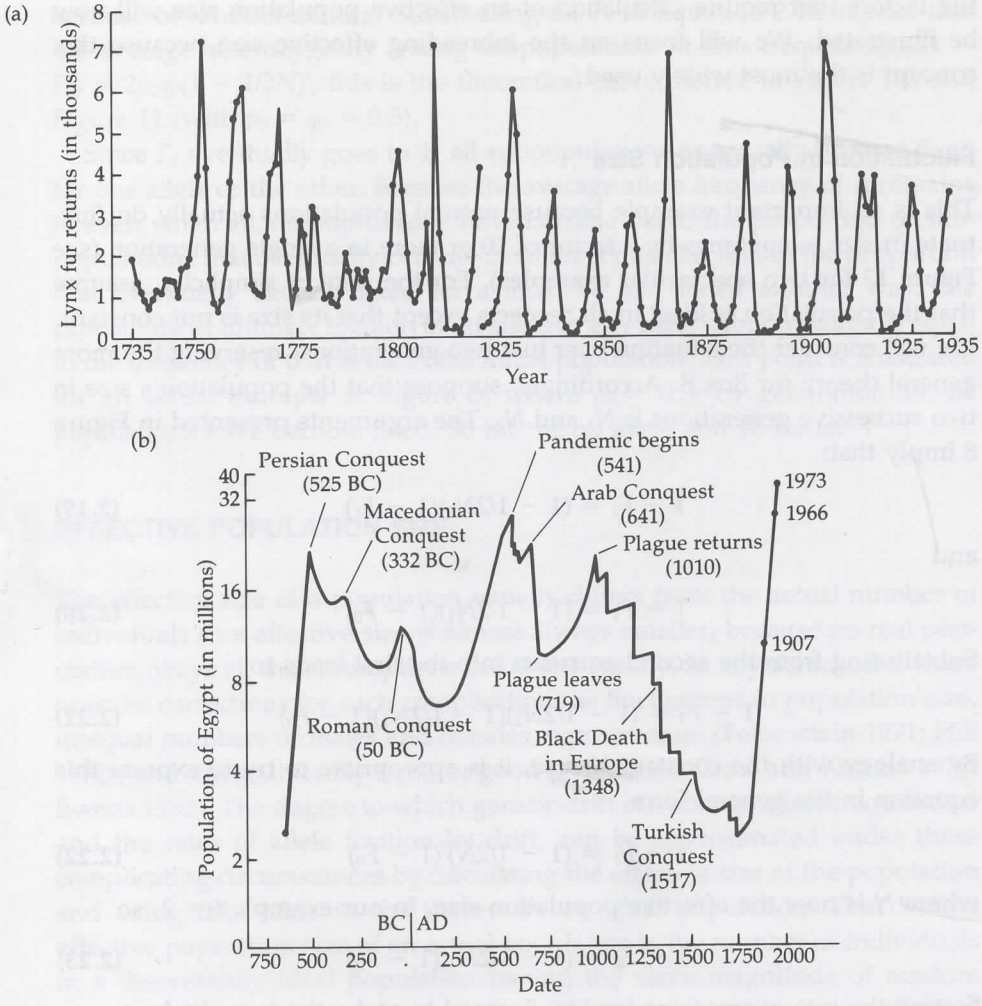
Setting the two expressions for $1 - F_2$ equal to each other we obtain

$$(1 - 1/2N)^2 = (1 - 1/2N_1)(1 - 1/2N_2) \qquad \textbf{(2.24)}$$

from which $1/N = (1/2)(1/N_1 + 1/N_2)$ turns out to be an excellent approximation (see Box B for justification). In general,

$$\frac{1}{N_e} = \left(\frac{1}{t}\right)\left(\frac{1}{N_1} + \frac{1}{N_2} + \ldots + \frac{1}{N_t}\right) \qquad \textbf{(2.25)}$$

so the effective size N_e is the **harmonic mean** of the actual numbers—the reciprocal of the average of reciprocals. The harmonic mean tends to be dominated by the smallest terms. Suppose, for example, that $N_1 = 1000$, $N_2 = 10$, and $N_3 = 1000$ in a population that underwent a severe temporary reduction in size (a **bottleneck**) in generation 2. Then $1/N_e = (1/3)(1/1000 +$



1/10 + 1/1000) = 0.034, or N_e = 1/0.034 = 29.4. The average effective number over the three-generation period is only 29.4, whereas the average actual number is (1/3)(1000 + 10 + 1000) = 670. Population bottlenecks are thought to account for the very low levels of polymorphism found in extant populations of the elephant seal (Bonnell and Selander 1974) and the cheetah (O'Brien et al. 1985, 1987). A severe population bottleneck often occurs in nature when a small group of emigrants from an established subpopulation founds a new subpopulation; the accompanying random genetic drift is then known as a **founder effect** (see Holgate 1966; Nei et al. 1975; Chakraborty and Nei 1977; Neel and Thompson 1978).

◁**Figure 12. Fluctuations in population size. (a) Famous cycles of population density in the Canadian lynx (*Lynx canadensis*) as reflected by the number of lynx furs brought to trading posts of the Hudson's Bay Company by trappers. The record covers 200 years and is one of the longest population records known. Values for 1735–1820 are from Hudson's Bay Company records for an area that is now Manitoba, northern Ontario, western Quebec, and part of Saskatchewan; points for 1921–1934 are from the Mackenzie River District, which included the western part of the District of Mackenzie (in the modern Northwest Territories) and small parts of Alaska, the Yukon, British Columbia, and Alberta. Juxtaposing the graphs is justified because lynx cycles are synchronized throughout continental Canada, though the "troughs" prior to 1821 are not as low as later ones because the trapping areas in the first and last parts of the graph are different. Note the regular fluctuations in population density (by a factor of 50 or more) with a period of 8 to 10 years (average 9.6). Why lynx populations are cyclic is not known for certain. It is interesting that the lynx eats mainly rabbits — the snowshoe hare, *Lepus americanus* ("It lives on Rabbits, follows Rabbits, thinks Rabbits, tastes like Rabbits, increases with them, and on their failure dies of starvation in the unrabbited woods," says hare observer E. T. Seton) — and the rabbits cycle too, with about the same period. (Data and quote from Elton and Nicholson 1942.) (b) Historical demographer's reconstruction of the population history of Egypt. Note large fluctuations due to plagues and conquests, and the recent upsurge. Extreme fluctuations in population size such as those shown here are not atypical, and their relevance to population genetics is that the effective size of a fluctuating population is markedly reduced by recurrent episodes of small population number. (From Hollingsworth 1969.)**

Effective Population Size

B

The effective size of a nonideal population is the size of an ideal population showing the same rate of increase in F per generation as the nonideal population. In this box, we derive expressions for effective population size in two important cases: (1) when the actual size varies from generation to generation, and (2) when there are unequal numbers of males and females. The actual populations are assumed to be ideal in all respects except the ones specified.

a. Because $e^{-x} = 1 - x + x^2/2 - x^3/3 + \ldots$, we may write $e^{-x} = 1 - x$ if x^2 is small compared to x. Motivated by the expression

$$(1 - 1/2N)^2 = (1 - 1/2N_1)(1 - 1/2N_2)$$

for the average effective size of a population over two generations (see text), define $x = 1/2N$, $x_1 = 1/2N_1$, and $x_2 = 1/2N_2$; assume that the N's are large enough that x^2, x_1^2 and x_2^2 are much smaller than x, x_1 and x_2. Then the above expression is approximately

$$(e^{-x})^2 = (e^{-x_1})(e^{-x_2})$$

Show from this that $2x = x_1 + x_2$, or

$$1/N_e = (1/2)(1/N_1 + 1/N_2)$$

and hence after t generations,

$$1/N_e = (1/t)(1/N_1 + 1/N_2 + \ldots + 1/N_t)$$

Suppose a population of insects in four consecutive generations has effective sizes of 10, 10^2, 10^3 and 10^4. What is the average effective population size over this period?

b. In Figure 8 it may be noted that the inbreeding effective population size may be defined as the reciprocal of the probability that two uniting gametes come from the same parent. When there are separate sexes, of course, uniting gametes must come from different parents; in such a case, the effective size may be defined as the reciprocal of the probability that two uniting gametes come from the same grandparent. Because both sexes must contribute equally to offspring, the probability that uniting gametes come from a male grandparent is 1/4. Given that the gametes both come from a male, the probability that they come from the same male is $1/N_m$, where N_m is the number of males in the grandparental generation; $1/N_f$ is the corresponding probability for a female grandparent. Altogether, then,

$$1/N_e = 1/4N_m + 1/4N_f$$

or

$$N_e = 4N_mN_f/(N_m + N_f)$$

among the grandparents. What is the effective size of a population of 100 cows fertilized by artificial insemination with the sperm of four bulls? What would it be if there were 200 cows?

ANSWER **a.** Since $(e^{-x})^2 = e^{-2x} \approx 1 - 2x$ (for x small), and $e^{-x_1} \approx 1 - x_1$, $e^{-x_2} \approx 1 - x_2$, we have $1 - 2x = (1 - x_1)(1 - x_2) = 1 - x_1 - x_2 + x_1x_2 \approx 1 - x_1 - x_2$ for x_1x_2 small. Thus $2x = x_1 + x_2$. For the insect population, $1/N_e = (10^{-1} + 10^{-2} + 10^{-3} + 10^{-4})/4 = 0.02778$, so $N_e = 36$.

b. In the case of 4 bulls and 100 cows, $N_e = 15.385$; for 4 bulls and 200 cows, $N_e = 15.686$.

Unequal Numbers of Males and Females

A second important case in which the effective size of a nonideal population can readily be calculated concerns sexual populations in which the number of males and females is unequal. This inequality creates a peculiar sort of "bottleneck": because half of the alleles in any generation must come from each sex, any departure of the sex ratio from equality will enhance the opportunity for random genetic drift. This situation is important in wildlife management where, for many game animals (pheasants and deer come immediately to mind), the legal bag limit for males is much larger than that for females. Although some management goals are served by such hunting regulations (for example, the species involved are usually polygamous, so

one male can fertilize many females and overall actual population size can be maintained), it must be remembered that the resultant inequality in sex ratio reduces the effective population size. Specifically, if a sexual population consists of N_m males and N_f females, the actual size is

$$N_a = N_m + N_f \tag{2.26}$$

However, as shown in Box B, the effective population size is

$$N_e = \frac{4N_mN_f}{N_m + N_f} \tag{2.27}$$

To take a realistic example, if hunting is permitted to a level at which the number of surviving males is one-tenth the number of females, then the effective population size is a mere one-third of the actual number of individuals in the population. (To see this for yourself, set $N_m = 0.1N_f$ in the formulas for N_a and N_e.)

A related problem that was addressed by a previous example (page 79) is the effective population size for an X-linked gene. In this case, the variance effective population size is

$$N_e = \frac{9N_mN_f}{4N_m + 2N_f} \tag{2.28}$$

This can be justified by noting that the sampling variance for the X chromosomes from males is p_mq_m/N_m and the sampling variance for X chromosomes from females is $p_fq_f/2N_f$, where p_m and p_f are the allele frequencies in males and females respectively. The frequency of an X chromosome in the population is

$$p = (1/3)p_m + (2/3)p_f \tag{2.29}$$

and the sampling variance of p is

$$\text{Var}(p) = (1/9)(p_mq_m/N_m) + (4/9)(p_fq_f/2N_f) \tag{2.30}$$

At steady state, $p_m = p_f = p$, so pq can be factored out, giving

$$\text{Var}(p) = pq[(1/9)(1/N_m) + (4/9)(1/2N_f)] = pq/2[9N_mN_f/(4N_m + 2N_f)] \tag{2.31}$$

The term in the square brackets corresponds to the N_e in Equation 2.28, showing why this is a variance effective size: the binomial sampling variance in an ideal population is $pq/2(N_e)$.

Overlapping Generations

The simplest model of random genetic drift with overlapping generations was presented by Moran (1962). The model is one of a haploid population

of constant size, in which, during each interval of time, one randomly chosen individual gives birth and another randomly chosen individual dies. Moran showed that the probability of drawing a pair of alleles that are not identical by descent (H) declines at the rate of $2/N^2$ per unit time. Since time is scaled with one generation equal to N time units, the rate of decline of H is $N(2/N^2) = 2/N$. Comparing this with the rate of decline of H in a discrete generation haploid population, which is $1/N$ per generation, the effective population size in the Moran model is $N_e = N/2$. Felsenstein (1971) elaborated considerably on this problem, and the general result is that the effective population size is smaller when generations overlap.

The diploid sexual model with overlapping generations can be quite complex, since males and females can have different age-specific schedules of reproduction and death. Hill (1972) found that the effective size of a population with overlapping generations was equal to the size of a discrete-generation population with the same number of individuals entering the population each generation and the same variance in lifetime family size. Letting N be the number of individuals born during each time period, L the mean generation time (the mean age of parents when offspring are born), and σ_n^2 the variance in family size, Hill obtained

$$N_e = \frac{4NL}{\sigma_n^2 + 2} \tag{2.32}$$

A larger variance in family size results in a smaller effective population size, for reasons detailed in the following section.

Non-Poisson Distribution of Fecundity

There is an implicit assumption in the Wright-Fisher model that the number of gametes that individuals contribute to the next generation follows a Poisson distribution. An important property of the Poisson distribution is that the mean equals the variance. Any departures from this distribution require recalculation of the effective population size. For example, if all genotypes produce exactly the same number of offspring, then the distribution of fecundity is too uniform. To take another example, if there is a strong skew in the fecundity distribution, then too few individuals produce a disproportionate fraction of the offspring. The assumption of a Poisson distribution in the Wright-Fisher model can be seen as follows (Crow and Kimura 1970). In a diploid population of size N, the probability that k gametes come from a particular parent follows the binomial distribution

$$\binom{2N}{k} (1/N)^k (1 - 1/N)^{2N-k} \tag{2.33}$$

This is true because the k gametes came from a particular individual with probability $1/N$, and the remaining $2N - k$ gametes came from the other individuals, whose frequencies sum to $1 - 1/N$. If the population is not growing, then on average each genotype contributes two successful gametes to the next generation, which we write as $\bar{k} = 2$. The variance in the number of gametes contributed is the variance of the above binomial, so

$$V_k = \bar{k}N(1/N)(1 - 1/N) = \bar{k}(1 - 1/N) \qquad \textbf{(2.34)}$$

For large N, the variance in the number of contributed gametes approaches $V_k = \bar{k}$, which is the same as the expected number of gametes contributed by each individual. The frequency distribution for the number of contributed gametes in the Wright-Fisher model is therefore the Poisson distribution.

In order to determine the inbreeding effective population number when the fecundities do not follow a Poisson distribution, Wright (1938) calculated the probability of autozygosity for non-Poisson distributions of gametes. In the special case of $\bar{k} = 2$ (i.e., a diploid population of constant size in which the mean contribution per individual is two gametes), the effective population size is

$$N_e = \frac{4N - 2}{V_k + 2} \qquad \textbf{(2.35)}$$

At one extreme, if every individual in the population contributes an identical number of gametes to the next generation, then $V_k = 0$, and therefore $N_e = 2N - 1$. In words, the effects of the random drift can be reduced by a factor of two by making sure that all individuals are represented equally in the successful gametes each generation. Ecological effects on the variance in fecundity can have a large effect on the effective population size, as documented in the *Euphydryas* butterfly by Mueller et al. (1985b). The effect of variance in fecundity is also important in zoo and wildlife management, where it is often desired to retain genetic variability in small populations. At the other extreme, if only one individual contributes all the gametes to the next generation (remember that self-fertilization is permitted), then the formula gives $N_e = 1$. A far-reaching conclusion can be drawn from these considerations of the effect of the variance in fecundity on effective population size. If natural selection is operating, then different individuals have different expected contributions to the next generation, and the effective population size must be smaller than the census size. Furthermore, the greater the variance in fecundity (or fitness), the smaller the effective population size. So we have revealed an apparent paradox — that as the strength of selection increases, so does the magnitude of the effects of random genetic drift.

Geographic Dispersion of Populations

The genetic consequences of population dispersion can be seen by contrasting the situation of a uniform distribution such as over a prairie versus a clumped distribution such as might occur on a group of islands. In this case the effective size depends on two quantities: (1) the number of breeding individuals per unit area, usually denoted by the symbol δ, and (2) the amount of dispersion between an individual's own birthplace and that of its offspring. The latter quantity, usually denoted by the symbol σ^2, is called the *one-way variance of distance between birth and breeding sites*. If dispersion follows a normal bell-shaped curve in both dimensions, then σ^2 has the following biological interpretation: 39% of all individuals have their offspring within a circle of radius σ centered at their own birthplace; 87% have their offspring within a circle of radius 2σ; and 99% have their offspring within a circle of radius 3σ. In terms of δ and σ, the effective size of the population (often called the **neighborhood size**) is given by

$$N_e = 4\pi\delta\sigma^2 \qquad (2.36)$$

where $\pi = 3.14159$.

Equation 2.36 can be applied to data on the abundant deer mouse *Peromyscus maniculatus*. In a large area in southern Michigan, Dice and Howard (1951) estimated the density of breeding individuals to be between $\langle\delta\rangle = 5$/ha and $\langle\delta\rangle = 7$/ha (where 1 ha = 1 hectare = 10,000 m^2, or about 2.5 acres). By following the movement of marked animals from birth to breeding sites, they estimated $\langle\sigma^2\rangle = 1.3$ ha (excluding woodlands, which this organism avoids). Applying Equation 2.36, the neighborhood size is estimated to be between $\langle N_e\rangle = 4(3.14159)(5)(1.3) = 82$ and $\langle N_e\rangle = 4(3.14159)(7)(1.3) = 114$. The neighborhood size is surprisingly small for such an abundant animal, but this result is consistent with direct measures of population structure using protein electrophoresis (Avise et al. 1979).

GENEALOGY AND THE COALESCENT, OR, ALL ABOUT EVE

A useful way to think about the Wright-Fisher model is to consider the genealogy of the members of a population in a pure drift model (Kingman 1980). This perspective is even more powerful when mutations are allowed, but we will ignore mutation in order to illustrate the principle. The idea is to consider a population that has been undergoing random genetic drift according to the Wright-Fisher sampling process, and to trace backwards in time to the ancestors of the individuals in the current population. The simplest example is a haploid population with the complete genealogy known.

Although each of the $2N$ individuals in the population at time t had parents in the population in generation $t - 1$, not all individuals in generation $t - 1$ produced offspring. Indeed, for any pair of individuals in the population, it is possible to trace their lineage back to a common ancestor. Similarly, any set of r individuals can be traced back to a common ancestor. In fact, at any given time the *entire population* can be traced back to a common ancestor, and the genealogy of the population is said to **coalesce** (Figure 13).

To examine the concept of coalescence, let us pick a set of r members of the population in generation t. Let ξ_n be the number of individuals in generation $t - n$ who are ancestors of at least one of the r individuals in generation t. A little reflection will justify the following:

$$r = \xi_0 \geq \xi_1 \geq \xi_2 \geq \xi_3 \geq \ldots \geq 1$$

The ξ's are independent of each other, and the transitions from ξ_n to ξ_{n+1} form a Markov chain with transition probability

$$T_{ij} = \Pr(\xi_{n+1} = j \mid \xi_n = i) \qquad \textbf{(2.37)}$$

where the probability is identical to the chance of observing j empty boxes when i balls are distributed randomly among $2N$ boxes. The probability can be written out explicitly, and the important feature of this Markov chain is

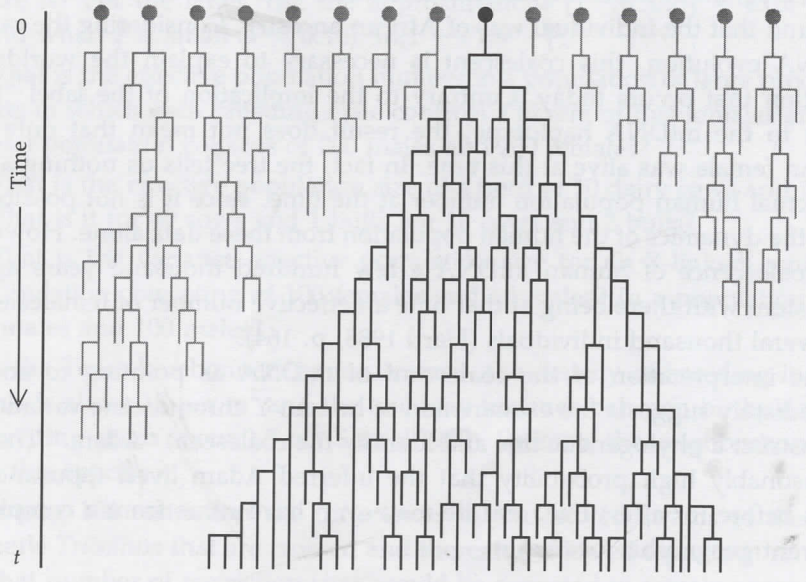


Figure 13. The coalescent, showing that many lineages go extinct (indicated in gray) and *all* alleles present in the population at time t trace back through the black tree to one common ancestral allele at some time in the past.

that powers of the transition probability matrix yield information about the genealogical relationships among the individuals in the population. For example, with respect to the common ancestry of two individuals ($r = 2$), Kingman (1980) showed that the probability of two individuals in generation t having a single common ancestor in generation $t - n$ is

$$T_{21}^{n} = 1 - \lambda^{n} \tag{2.38}$$

where $\lambda = 1 - 1/(2N)$. The concept of genealogies of genes and tracing backwards in time to the coalescent will be referred to again when the theory of neutral mutations is introduced. By comparing inferred genealogies with those predicted from the neutral theory, inferences can be made about the history of population genetic forces acting on polymorphisms.

The concept of coalescence of a genealogy has recently appeared in the popular press in the interpretation of human mtDNA variation (for the scientific account, see Cann et al. 1987). The mtDNA restriction site variation among 147 humans has been studied in great detail and used to reconstruct phylogenetic trees indicating the pattern of relatedness of mtDNA among extant humans. Various assumptions are made in constructing such trees, which are discussed in Chapter 7, but for now the trees will be accepted as providing an approximate time scale for the lineages since their common ancestor. The phylogenetic tree of human mtDNA was extrapolated to infer that the variation coalesced into one individual who lived about 200,000 years ago, and that the individual was of African ancestry. Considering the rate of mtDNA evolution, this coalescent is necessary to explain the worldwide variation that occurs today. Contrary to the implication of the label "Eve" given to the mtDNA haplotype, the result does not mean that only one human female was alive at this time. In fact, the tree tells us nothing about the actual human population number at the time, since it is not possible to infer the dynamics of the human population from these data alone. However, the coalescence of human mtDNA a few hundred thousand years ago is consistent with there being at that time an effective number of females equal to several thousand individuals (Hartl 1988, p. 164).

The interpretation of the coalescent of mtDNA as pointing to an Eve immediately suggests the examination of human Y chromosome variation to reconstruct a phylogenetic tree and identify the coalescent "Adam." There is a reasonably high probability that the inferred Adam lived thousands of years before (or after) Eve, and he may even have come from a completely different geographic location.

PROBLEMS

1 A gene in one individual in a population of 24 barn cats undergoes mutation to a new neutral allele. What is the probability that the allele eventually becomes fixed? What is the probability that it eventually becomes lost? What are the answers if the mutant gene is X-linked and the population consists of equal numbers of males and females?

2 If an isolated population of annual alpine plants decreases in heterozygosity by half every 50 years because of random genetic drift, what is its effective population size?

3 Remote Pitcairn Island in the South Pacific was settled in 1789 by Fletcher Christian and eight fellow mutineers from HMS *Bounty*, along with a small number of Polynesian women. Although many of their descendants have left the island in the intervening years, there has been essentially no immigration. Assuming an effective size of 20 in each of the eight generations since the island's settlement, what value of F_{ST} would be expected in today's population from random genetic drift?

4 In a population of effective size $N = 50$, how many generations are required for random genetic drift to double the value of the fixation index F from 0.01 to 0.02? From 0.05 to 0.10? Assuming that F is small, how many generations are required to double the value of F in a population of effective size N? For the latter, use the approximations $[1 - (N/2)]^t \approx \exp(-t/2N)$ and when F is small ($F < 0.10$), $\ln(1 - F) \approx -F$.

5 What is the effective population number in a population of large predatory cats in which each breeding male controls a harem of five females and the total population consists of 200 males and 200 females?

6 What is the effective population size of a herd of 10 dairy cows and 1 bull? What is it for 40 cows and 1 bull? For 10 cows and 2 bulls?

7 What is the variance effective population size for an X-linked gene in a population consisting of 100 females and 10 males? In a population of 10 females and 100 males?

8 In the Texas bluebonnet *Lupinus texensis*, the density of breeding individuals is about 15 per m^2 and the one-way variance between birthplace and breeding site is about 0.5 m^2 (Schaal 1980). Estimate the neighborhood size in this species.

9 Among 100 restriction site differences in two inbred strains of the flour beetle *Tribolium* that are crossed and thereafter allowed to mate at random, what number of restriction sites would be expected to remain segregating after 10 generations, assuming an effective population size of 80 individuals? How many would be expected to remain unfixed after 50 generations?

10 In a geometric progression $x, x^2, x^3, \ldots, x^n$, the sum of the first n terms equals $x(1 - x^n)/(1 - x)$. Use this to determine the effective population size in generation n of a population that grows according to the progression 2, 4, 8, 16, 32, 64, 128, 256, . . . , 2^n. What is the limiting effective size?

11 What is the effective size of an otherwise ideal population in which the distribution of successful gametes per individual has a mean of two and a variance of four?

12 With reference to Equation 2.35:

a. What is the effective population size of a diploid population in which each individual contributes exactly two successful gametes to the next generation?

b. Why does random genetic drift occur at all in this situation?

c. What would you judge is the effective population size of a haploid asexual population in which each individual contributes exactly one successful offspring to the next generation?

13 In a population of constant effective size 50, what is the probability that two random individuals shared a common ancestor 100 generations ago?

14 In a population that is growing in size, does Equation 2.38 overestimate or underestimate the probability that two random individuals share a common ancestor in generation $t - n$?

15 Use Equation 2.38 to show that the probability of two individuals sharing a single common ancestor N generations ago is approximately 0.39 in a population of constant effective size N. Bear in mind that $(1 - x)^y \approx \exp(-xy)$ if xy is not too large. Verify the result numerically for a population with $N = 50$.

16 Show that random genetic drift requires an average of $t = 2N\ln x$ generations to reduce the heterozygosity from H_0 to H_0/x.

17 Use Equation 2.18 to show that approximately $2N$ generations of random genetic drift are required to reduce the number of segregating genes by a factor of e (e = 2.71828. . .), given initial allele frequencies close to 0.5.

18 For a single copy of a neutral allele in a population of effective size N_e and actual size N_a, it can be shown that, if the allele is destined to be fixed, the average number of generations until fixation is approximately $4N_e$; and if the allele is destined to be lost, the average number of generations until loss is approximately $(2N_e/N_a)\ln(2N_a)$. Make a table of expected times to fixation and loss for the following populations:

	Population									
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
N_a	100	200	400	800	1600	3200	6400	12,800	25,600	51,200
N_e	80	160	320	640	1280	2560	5120	10,240	20,480	40,960

19 For an allele with frequency p in a population of effective size N_e, if the allele is destined to be fixed, the average number of generations until fixation is approximately

$$t_1 = -(1/p)[4N_e(1 - p)\ln(1 - p)]$$

and if the allele is destined to be lost, the average number of generations until loss is approximately

$$t_0 = -4N_e[p/(1 - p)]\ln(p)$$

For $p = 0.5$, one might expect $t_1 = t_0$. Show that this is indeed the case. Calculate the common value of t_1 and t_0 for $p = 0.5$.

20 Random genetic drift in a population of size 1 corresponds to repeated self-fertilization. Consider a population established from heterozygous *Aa* monoecious plants and maintained by self-fertilization by choosing one random progeny from each parental plant, and let the vector $\mathbf{x}_n = ([AA]_n, [Aa]_n, [aa]_n)$ be the frequencies of the genotypes after n generations. What is the transition matrix **T** for the number of *A* alleles (0, 1, or 2) in each individual plant? Use **T** to calculate $\mathbf{x}_1$, $\mathbf{x}_2$, and $\mathbf{x}_3$, and try to infer a general expression for $\mathbf{x}_n$. Note that $\mathbf{x}_0 = (0, 1, 0)$.

21 Consider the populations of size 1 in Problem 20 and suppose that the genotype *aa* is lethal owing to a chlorophyll deficiency. In this case a viable plant can have only two states (genotypes), corresponding to the presence of one copy of *A* (genotype *Aa*) or two copies (genotype *aa*). What is the transition matrix **T** in this case. Letting $\mathbf{x}_n = ([Aa]_n, [AA]_n)$ be the vector of genotype frequencies in generation n, calculate $\mathbf{x}_1$, $\mathbf{x}_2$, and $\mathbf{x}_3$, and try to infer a general expression for $\mathbf{x}_n$, assuming that $\mathbf{x}_0 = (1, 0)$.

Chapter Three

Mutation and the Neutral Theory

ORIGIN OF GENETIC VARIATION

Mutation is the ultimate source of genetic variation. This is an important statement, but hardly a profound one. New genetic variation is created by changes in the genetic material, and heritable changes in the genetic material are, by definition, mutations. The word *mutation* is used here in its widest sense to mean all genetic changes, including visible chromosome abnormalities such as inversions and translocations, and also spontaneous polyploidy. **Polyploidy** refers to organisms or species having multiple sets of chromosomes. Stebbins (1976) has shown that polyploidy is of immense importance in the origin of new species in higher plants. However, the vast majority of genetic variation within virtually all natural populations consists of changes in DNA that are invisible through the microscope. At the molecular level, one class of mutations results from the replacement of one pair of nucleotides with another, which includes the mutations responsible for allozymes. Another class consists of deletions or duplications of nucleotide sequences. Still another class results from the transposition of DNA sequences from one location in the genome to another. Transposition mutations often result from the movement of DNA sequences called transposable elements, many of which contain genes that promote their own transposition. Mutations resulting from transposable elements constitute an important source of genetic variation, and the implications of transposable elements in population genetics are discussed in Chapter 7.

Because spontaneous mutation rates are typically rather small (on the order of 10^{-4} to 10^{-6} mutations per gene per generation), the tendency for allele frequencies to change as a result of recurrent mutation (**mutation pressure**) is negligible over the course of a few generations. On the other hand, the cumulative effects of mutation over long periods of time can become appreciable. A useful model for thinking about the problem is the Hardy-Weinberg model of Chapter 1, but with mutation allowed. For the moment we focus on mutations that are selectively neutral — mutations that have so little effect on the ability of the organism to survive and reproduce that natural selection does not appreciably influence their frequency.

Classes of Mutation

Although most of classical population genetic theory can be developed with a very simple concept of mutation, it is worthwhile to begin with some molecular aspects of mutation. Traditionally, mutation meant simply a spontaneous change from allele *A* to allele *a*, and the **rate** of mutation was the fraction of gametes that underwent mutation in one generation. Considering mutation at the molecular level, it is important to distinguish among mutation rates. A mutation rate might be expressed as mutations per nucleotide, as mutations per gene, or as mutations per gamete.

The simplest kind of mutation is a **base substitution**, in which one nucleotide is replaced by another. If the mutation is from one purine to another (adenine to guanine or vice versa), or from one pyrimidine to another (cytosine to thymine or vice versa), then the mutation is a **transition**. All other types of nucleotide substitutions are **transversions**. There are twice as many possible transversions as transitions. If a base substitution occurs in a codon within a protein-coding region, this may result in an amino acid change in the primary structure of the protein (see the genetic code, Table 1). If a nucleotide substitution causes no change in the protein product of the gene, such a mutation is known as a **synonymous substitution** (or a silent mutation). It is natural to suppose that, since the gene product is not affected, synonymous substitutions have no effect on fitness (they are **selectively neutral**). However, situations can occur in which silent mutations make a difference in the expression of a gene. For example, the secondary structure (pattern of intramolecular folding) of DNA can be very important in determining the accessibility of RNA polymerase and other proteins involved in gene expression. Examples of putative selective effects attributed to silent mutations will be encountered in Chapter 7.

If a base substitution results in the replacement of an amino acid in the protein product, it is a **nonsynonymous** (or missense) mutation. Allozymes

Table 1. The genetic code.[a]

First nucleotide in codon	Second nucleotide in codon and translation product							
	U	Translation product	C	Translation product	A	Translation product	G	Translation product
U	UUU	phe(F)	UCU	ser(S)	UAU	tyr(Y)	UGU	cys(C)
	UUC	phe(F)	UCC	ser(S)	UAC	tyr(Y)	UGC	cys(C)
	UUA	leu(L)	UCA	ser(S)	UAA	*ochre*(stop)	UGA	*opal*(stop)
	UUG	leu(L)	UCG	ser(S)	UAG	*amber*(stop)	UGG	trp(W)
C	CUU	leu(L)	CCU	pro(P)	CAU	his(H)	CGU	arg(R)
	CUC	leu(L)	CCC	pro(P)	CAC	his(H)	CGC	arg(R)
	CUA	leu(L)	CCA	pro(P)	CAA	gln(Q)	CGA	arg(R)
	CUG	leu(L)	CCG	pro(P)	CAG	gln(Q)	CGG	arg(R)
A	AUU	ile(I)	ACU	thr(T)	AAU	asn(N)	AGU	ser(S)
	AUC	ile(I)	ACC	thr(T)	AAC	asn(N)	AGC	ser(S)
	AUA	ile(I)	ACA	thr(T)	AAA	lys(K)	AGA	arg(R)
	AUG*	met(M)	ACG	thr(T)	AAG	lys(K)	AGG	arg(R)
G	GUU	val(V)	GCU	ala(A)	GAU	asp(D)	GGU	gly(G)
	GUC	val(V)	GCC	ala(A)	GAC	asp(D)	GGC	gly(G)
	GUA	val(V)	GCA	ala(A)	GAA	glu(E)	GGA	gly(G)
	GUG	val(V)	GCG	ala(A)	GAG	glu(E)	GGG	gly(G)

[a] The asterisk indicates the "start" codon for protein synthesis; although methionine (met) is always the first amino acid in a polypeptide chain, it is frequently cleaved off the completed chain by enzymes. There are three "stop" codons, which signal the termination of translation; these are conventionally denoted *ochre, amber,* and *opal.* Two forms of abbreviation of amino acids are given in the table: one, a three-letter abbreviation and the other, a single-letter abbreviation (the latter is particularly useful for giving the amino acid sequence of long polypeptide chains). Three-letter abbreviations: ala, alanine; arg, arginine; asn, asparagine; asp, aspartic acid; cys, cysteine; glu, glutamic acid; gln, glutamine; gly, glycine; his, histidine; ile, isoleucine; leu, leucine; lys, lysine; met, methionine; phe, phenylalanine; pro, proline; ser, serine; thr, threonine; trp, tryptophan; tyr, tyrosine; val, valine.

are a result of nonsynonymous mutations. Some nonsynonymous mutations can have profound effects on the physiology of an organism, because single amino acid replacements can alter the function of protein products. Many human genetic disorders are due to single amino acid changes, including, for example, phenylketonuria, albinism, G6PD deficiency, and sickle-cell anemia. If a base substitution results in the generation of a new stop codon, the mutation almost always destroys the function of the gene product. Such a termination mutation is a **nonsense** mutation.

Mutations can also result in gain or loss of genetic material. If DNA is lost, for example from ionizing radiation, the mutation is a **deletion**, and the chromosome that results has a **deficiency**. If a deletion is very small, only one base or two, then it can result in shifting the translational reading frame of the messenger RNA. All codons downstream from a frame-shifting mutation will be "out of sync," and hence incorrect amino acids will occur downstream of the reading frame-shift in the final product. These are **frame-shift mutations**, and usually result in nonfunctional gene products. Frame-shift mutations can also be caused by the **insertion** of a base or two — or any number as long as the number of bases in the insertion or deletion is not a multiple of three.

Genetic material can also be rearranged. When a broken part of a chromosome flips end for end and rejoins in a reversed orientation, the result is an **inversion** (Figure 1). Inversions may be harmless as long as no genetic material is lost and no important genes are disrupted at the breakpoints of the inversion. If an important gene is disrupted, then an inversion might be lethal, and the subsequent population dynamics of the mutation would be just like that of any other lethal mutation.

The best known inversions are those in the third chromosome of *Drosophila pseudoobscura*. More than 20 different inversions are known in this species, and they are evidently maintained in populations over long periods. The case for natural selection maintaining the inversions was most vigorously made by Theodosius Dobzhansky, who did extensive studies in both field and laboratory populations (Chapter 4).

Inversions have another property that make them unique. The products of recombination from individuals that are heterozygous for an inversion can be abnormal, and either fail to end up in gametes (as in *Drosophila*) or lead to aneuploid gametes. In *Drosophila*, only the nonrecombinant gametes are recovered, and inversions have the apparent effect of suppressing recombination. The suppression has the effect of tying together the alleles in the inverted chromosome, resulting in a sort of "supergene." The term "coadapted gene complex" was used by Dobzhansky and others to refer to the alleles linked in inversions that attained high frequency in populations.

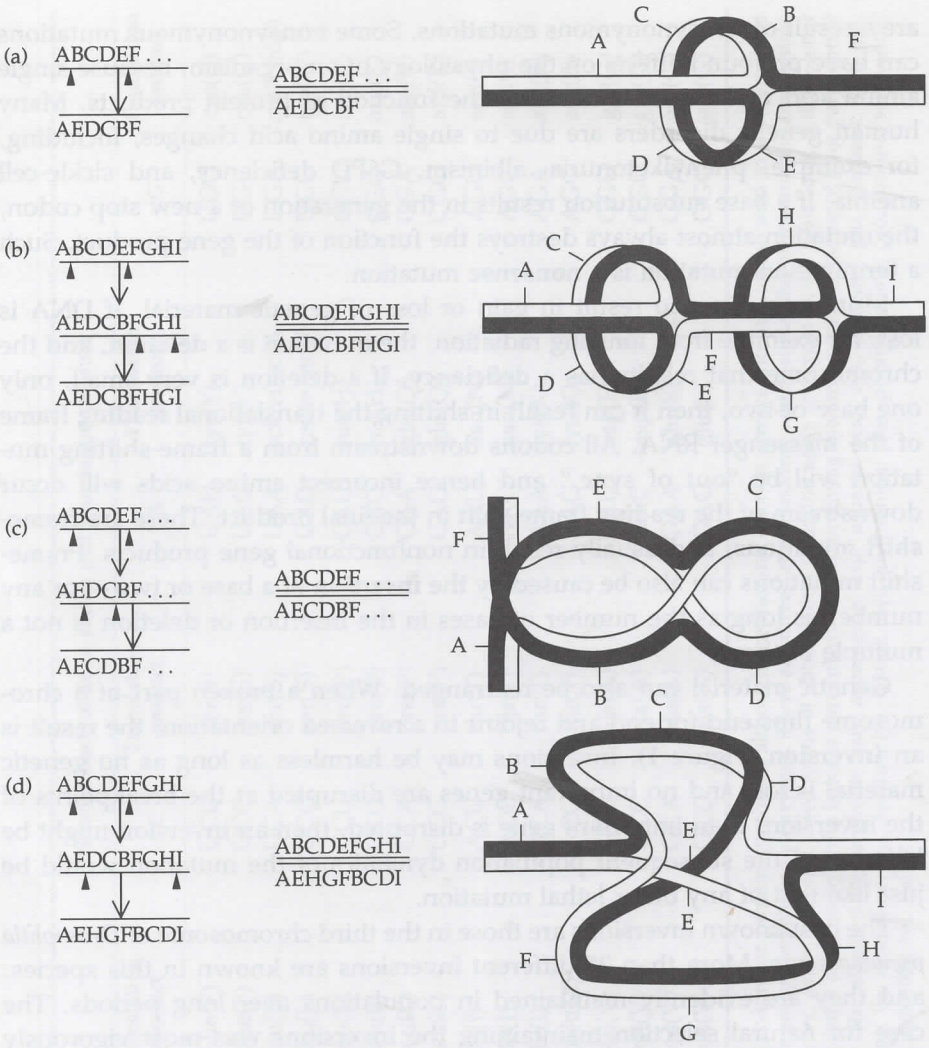


Figure 1. Chromosome pairing in the salivary gland cells of *Drosophila* heterozygous for inversions. (a) A single inversion; (b) two independent inversions; (c) two included inversions; (d) overlapping inversions. (From Dobzhansky 1970.)

Translocations are another form of rearrangement of genetic material that result from an exchange of material between nonhomologous chromosomes. Translocations have received particular attention from investigators studying population cytogenetics to learn about the evolution of the karyotype. Many organisms (particularly mice) contain races that differ in chromosome number due to Robertsonian translocations. Robertsonian translocations result from fusion of acrocentric chromosomes at the centromere, turning a pair of nonhomologous acrocentric chromosomes into a single metacentric chromosome.

In the mouse examples, both races have the same genetic material, but the karyotypes are arranged differently. The frequency of Robertsonian translocations and other types of karyotypic change differs greatly among species, suggesting that some organisms have a greater propensity for tolerating such events.

Mutations due to **gene conversion** come from misalignment of DNA. This phenomenon, illustrated in Figure 2, is especially important in the concerted evolution of tandemly repeated clusters of related genes, known as multigene families. Gene conversion is often associated with ordinary meiotic recombination. Classical crossing over takes place when there are two pairs of homologous sister chromatids, and each chromatid consists of a duplex molecule of DNA. If homologous alleles differ in nucleotide sequence, then after recombination, there may be a heteroduplex region of DNA that has a site or group of sites that are mispaired. In normal crossing over, these would separate, and the gametes would contain equal numbers of the two parental allelic types. On the other hand, if DNA repair takes place to correct the mismatch in the heteroduplex, then one or the other allele may be present in excess in the gametes. In essence, the mismatch repair converts one allele to the other. In some fungi, such as *Ascomycetes* and *Neurospora*, it is possible to recover all of the meiotic products and to detect gene conversion events directly. Gene conversion can take place with or without recombination of flanking genes: all that is necessary is that heteroduplex DNA be formed in the critical region, either between actual alleles or between nonidentical but similar genes. For any two alleles, *A* and *a*, conversions from *A* to *a* or from *a* to *A* may be equally likely. In some cases gene conversion is biased and one allele is preferentially converted. Biased gene conversion can have an impact on gene frequencies, particularly in multigene families, where it can tend to homogenize the family. Details of this process will be presented in Chapter 7.

Many mutations can be recognized and corrected by one of a number of DNA repair mechanisms. For example, UV radiation can induce covalent bonds between adjacent thymine bases and result in the formation of **thymine dimers**. The DNA replication machinery is often disrupted by thymine dimers, so they could be potentially lethal. Fortunately, thymine dimers are also immediately excised and replaced with normal nucleotides. The existence of mechanisms of repairing thymine dimers is inferred from mutations that lead to loss of these functions. For example, mutations such as *uvr* in *Escherichia coli*, and mutations that cause xeroderma pigmentosum in humans lead to unusually high sensitivity to UV radiation. In the case of human xeroderma pigmentosum, individuals must stay out of the sun lest they develop skin cancers from exposure to ultraviolet light. Another extraordinary DNA repair mechanism in *E. coli* is called SOS repair, in reference to

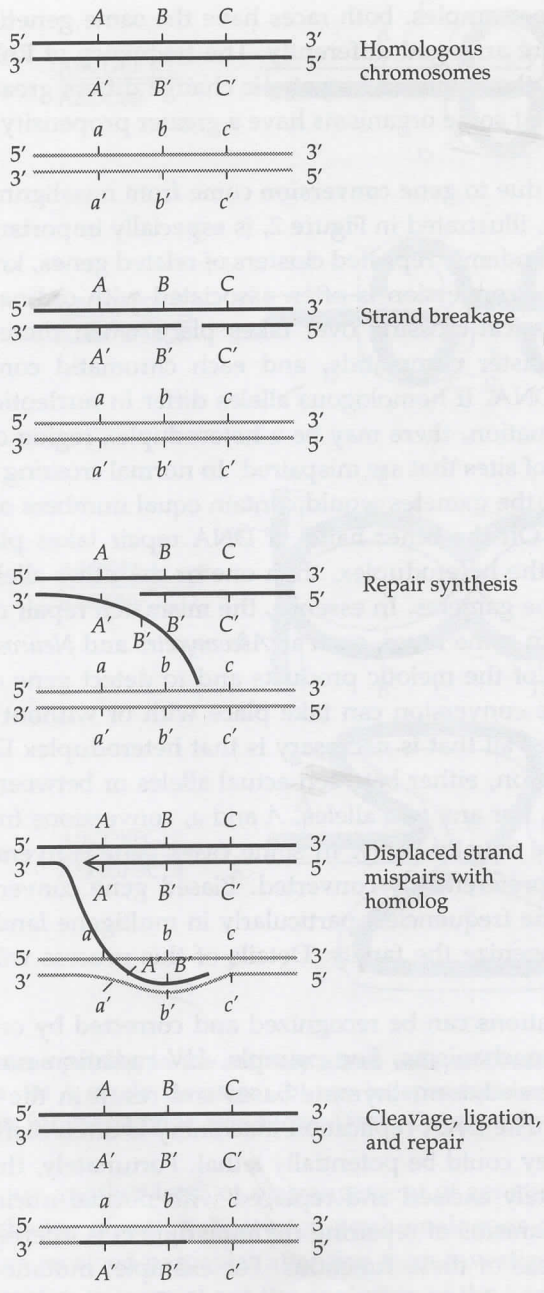


Figure 2. Gene conversion between two homologous DNA sites. The letters *A*, *B* and *C* represent three genes, each with two alleles in this heterozygote. The primes indicate the complementary strand of DNA. Single-strand breakage followed by mispairing of strands can ultimately result in the replacement (conversion) of one sequence by another. (After Darnell, Lodish and Baltimore 1986.)

the emergency nature of the process. In SOS repair, enzymes are induced that cleave damaged bases and repair the gaps. One interesting evolutionary feature of SOS repair is that the SOS polymerase makes frequent errors, so that SOS repair results in clusters of mutations when repair events occur. This might sound like an inefficient system, but the single lesion that is repaired might have been fatal, whereas the induced mutations might be minor. The evolutionary significance of DNA repair is that different classes of mutations have different likelihoods of being repairable: thymine dimers are easily recognized and easily repaired, but there is no mechanism to recognize and repair an inversion. Moreover, the observed rate of mutation must also depend on the rate and efficiency of repair.

There are also differences in mutations that affect their likelihood of being advantageous or deleterious. For example, most nonsense mutations are deleterious because they effectively eliminate the function of a gene. Silent substitutions are often neutral or nearly so. Missense mutations may have negligible effects, or if amino acid replacements occur in important parts of a protein, they can be disastrous. For example, changes in the active site of an enzyme can result in loss of function. The β^S globin allele is another example, in this case affecting the ability of the molecule to combine with oxygen. Since organisms have been evolving for a long time and consist of complex, highly integrated systems at several levels of organization (biochemical, physiological, nervous), it is expected that most random changes would be deleterious. However, one class of mutations may have a somewhat higher likelihood of producing evolutionary novelty than others. These are the rearrangements of genetic material — for example, the exon-shuffling types of rearrangements that combine different protein-folding domains (Chapter 7). Gene rearrangements in somatic cells are instrumental in generating the phenomenal antibody diversity of the immune system.

ESTIMATION OF MUTATION RATE

Since one of the basic parameters in population genetics is the mutation rate, estimating the parameter is necessary for interpreting and evaluating many models. Just as there are many classes of mutations, there are many ways to estimate the mutation rate. Descriptions of a few of the approaches follow.

Spontaneous Lethal Mutation Rate in *Drosophila*

Timofeeff-Ressovsky (1935) made use of features of *Drosophila* genetics to obtain estimates of X-linked lethal recessive mutations. Figure 3 shows the method. The *CIB* chromosome bears an inversion (formerly called C factors)

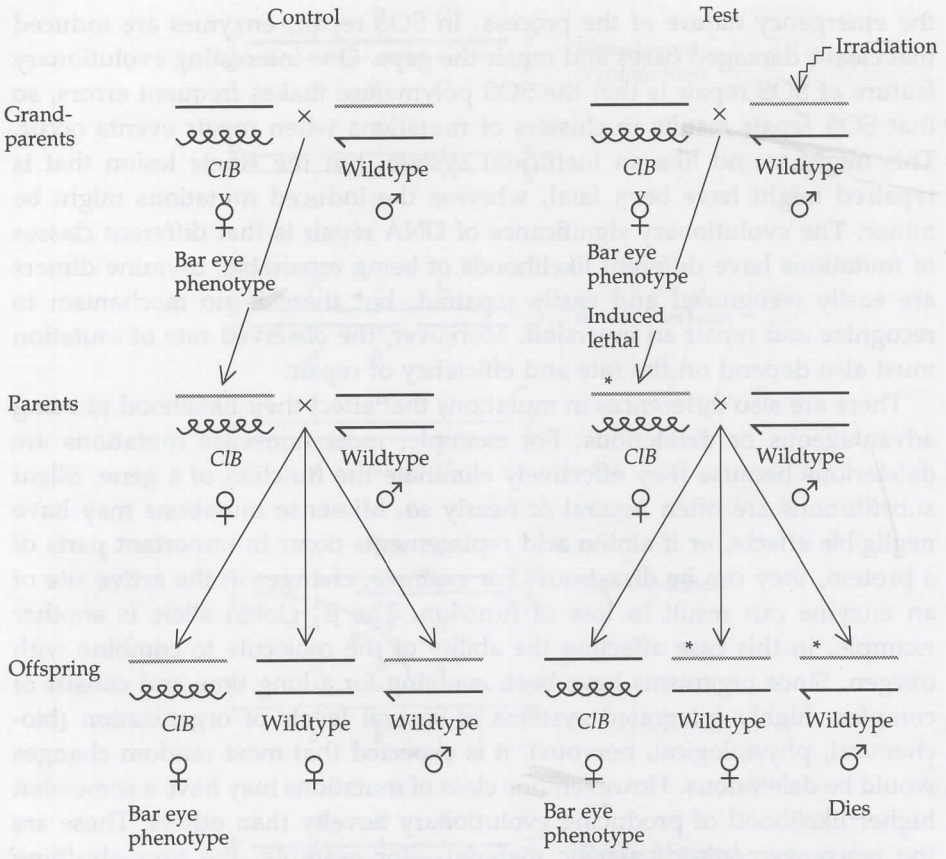


Figure 3. The crossing scheme used by Timofeeff-Ressovsky to reveal X-linked mutations. The expected ratio of the offspring in the control experiment is 1 *ClB* female:1 wildtype female:1 wildtype male. If an X-linked lethal were induced by radiation, the outcome in the offspring generation would be 1 *ClB* female:1 wildtype female:0 wildtype males. If the mutation were semiviable, a few wildtype male offspring may be observed. This design was originally used by Muller (1927) to detect the induction of X-linked lethals, and Timofeeff-Ressovsky pioneered the use of this design to quantify mutations of small effect.

that suppresses recombination by a mechanism described in the previous section. Multiple inverted chromosomes are known as **balancers**, and they play an important role in *Drosophila* genetics. The *ClB* chromosome also bears a recessive lethal (denoted *l*), and the *B* represents the dominant mutation *Bar* for narrow eyes. Males to be tested are crossed to females that are heterozygous for the *ClB* chromosome. The *Bar*-eyed daughters from this cross have an X chromosome from their father and the *ClB* chromosome from their mother. When these females are outcrossed to wildtype males,

the sons receive either the *ClB* X chromosome, and die at an early stage in development, or the grandfather's X chromosome. If the grandfather's X had mutated to a new recessive lethal, then no sons would be recovered from the testcross. If the X chromosome had mutated to a deleterious but not lethal allele, then some males would be recovered, but they would be less frequent than the female progeny. Figure 4 shows the graphs of sex ratios of grand-offspring of males treated with X-rays and untreated control males. Timofeeff-Ressovsky's experiment was not large enough to recover any spontaneous X-linked lethals in the controls, but the X-rays induced X-linked lethals in 13% of the experimental crosses. A balancer chromosome such as *ClB* is not necessary to observe a shift in sex-ratio after radiation-induced mutation. Because males only have one copy of the X chromosome, new X-linked lethal mutations result in fewer males in the population until the lethal X chromosomes are eliminated. For example, a shift in sex ratio toward females was observed in the natural *Drosophila* populations on Bikini atoll after the atomic bomb tests.

Wallace (1958) designed a method to detect mutations on the second chromosome of *Drosophila melanogaster* that are lethal when homozygous (Figure 5). The scheme makes use of a balancer chromosome containing *Curly* wing (*Cy*) and *Lobe* eye (*L*) markers that are lethal when homozygous, as well as multiple overlapping inversions to prevent the recovery of recombinant gametes. The crossing pattern in Figure 5 shows how to generate stocks

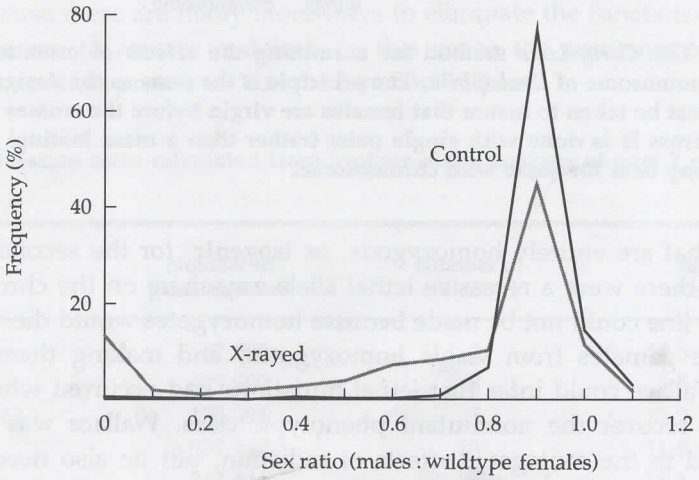


Figure 4. The results of Timofeeff-Ressovsky's (1935) X-linked lethal mutation study. A sex ratio of zero means an X-linked lethal was generated. Notice that the X-rayed flies (graphed in gray) also produced intermediate sex ratios, indicating deleterious, but not lethal, mutations.

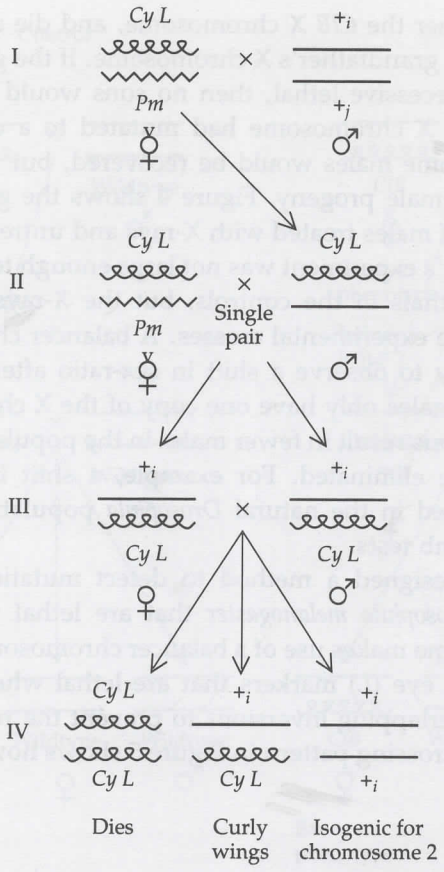


Figure 5. The *Curly-Lobe* method for examining the effects of mutations on the second chromosome of *Drosophila*. The principle is the same as the design in Figure 3. Care must be taken to insure that females are virgin before the crosses are set up, and that cross II is done with single pairs (rather than a mass mating) to be sure that progeny bear the same wild chromosome.

of flies that are entirely homozygous, or **isogenic**, for the second chromosome. If there were a recessive lethal allele anywhere on the chromosome, then this line could not be made because homozygotes would die. By recovering the gametes from viable homozygotes, and making them isogenic again, Wallace could infer that lethal mutations had occurred whenever he failed to recover the nonmutant phenotypic class. Wallace was primarily interested in the mutagenic effects of radiation, but he also needed to do unirradiated control experiments. Out of 5791 control second chromosomes tested in this way, 41 were found to be lethal. Using another balancer for the third chromosome, Wallace found that 38 out of 5864 third chromosomes were lethal after one generation. The combined average mutation rate was

79/11,655 = 0.0068 per second (or per third) chromosome per generation. Since there are about 2000 visible bands in the polytene second chromosome, which correspond approximately to the number of genes, the estimated mutation rate per gene per generation is $0.0068/2000 = 3.5 \times 10^{-6}$. It is likely that not all genes can mutate to lethals, and the figure of 2000 genes per chromosome is quite rough, so the actual rate of mutation per gene may vary considerably.

Coat Color Mutations in Mice

The design of the experiments of Russell (1963) and Schlager and Dickie (1971) was quite direct, and although they were very labor intensive, they yielded important information about the rate of mutation in mammals. Schlager and Dickie used the five recessive coat color mutations *nonagouti*, *brown*, *albino*, *dilute*, and *leaden* to examine offspring from brother–sister matings of 28 inbred strains. Dominant mutations could be observed in the offspring, and recessive mutations were recovered by observing a recessive phenotype in F_1 mice from a cross between the recessive homozygous line and wildtype. The experiment took over six years to complete and involved more than seven million mice. Table 2 reports the rates for the mutations from wildtype to the recessive mutant allele, and the rates of mutation from the recessive to the dominant allele. The rate of mutation to the recessive allele averaged 11.2×10^{-6} per gene per generation, and the reverse mutation rate was 2.5×10^{-6} per gene per generation. Generally, the reverse mutation rate is lower because there are many more ways to eliminate the function of a gene than there are to reverse a defect. In this sense the recessive coat color mutations are interpreted as "loss of function" alleles.

Table 2. Mutation rates calculated from brother–sister matings of over 7 million mice.

Locus[a]	Number of gametes tested	Number of mutations	Mutation rate ($\times 10^{-6}$)
a^- (*albino*)	67,395	3	44.5
b^- (*brown*)	919,699	3	3.3
c^- (*nonagouti*)	150,391	5	33.2
d^- (*dilute*)	839,447	10	11.9
ln^- (*leaden*)	243,444	4	16.4
	2,220,367	25	11.2 (average)

[a] Mutation is from wildtype to the recessive allele shown.
(From Schlager and Dickie 1971.)

Mutation Rate of Allozymes in *Drosophila*

Voelker et al. (1980) reported the results of a large study designed to give estimates of the rates of mutation at seven allozyme genes in *Drosophila melanogaster*. One thousand lines of flies bearing a pair of balanced lethal chromosomes were tested by gel electrophoresis after being maintained in the lab for an average of 220 generations. The balanced lethal configuration avoids the effect of segregation, which would lead to the loss of an allele, and the alleles are maintained in a heterozygous state, except in the extremely unlikely event of two identical mutations. Because mutations can accumulate over generations, examination of 100 chromosomes after five generations should result in the same number of mutations as in a sample of 500 chromosomes after one generation. An *allele generation* is the product of genes sampled times the number of generations. A total of only four mobility variants and 12 null (no detectable enzyme activity) mutations were detected out of 3,111,598 allele generations (Table 3). The rate of mutation to a new mobility variant was 1.28×10^{-6} per gene per generation and the rate of mutation to a null allele was 3.86×10^{-6} per gene per generation.

Mutation Rate of Polypeptides in Humans

Humans are not an ideal research organism for most genetic studies because of their size, long generation time, and possibilities for legal entanglements. Occasionally, however, we do perform catastrophic experiments on ourselves, and are then compelled at least to try to learn something from the mistake. One such catastrophe was the detonation of nuclear bombs over the cities of Hiroshima and Nagasaki during the Second World War. Extensive studies of the genetic consequences of the radiation have been done including analyses of karyotypic changes, birth defects, and electrophoretic assays of protein variation. Neel et al. (1986) recently published the results of an electrophoretic study of survivors of the Hiroshima bombing (the word *hibakusha* means "survivor of the bomb" in Japanese). The study included individuals who were more than 2500 meters from the blast; they served as a control for the studies of those who received high radiation doses. The 36 polypeptides that were examined covered a wide range of functions, including hemoglobin, albumin, esterase, G6PD, and PGM. A total of 539,170 "locus tests" were done, where a locus test refers to knowledge of both parents and one allele from one offspring. After excluding cases of mistaken paternity, there were three "exceptional" offspring whose allele was different from any parental allele. From these data, Neel and coworkers derived an estimate of the mutation rate of 5.6×10^{-6} per gene per generation, with a 95% confidence interval of $1\text{–}16 \times 10^{-6}$. Carrying the calculations further, if

Table 3. Mutation rate of allozymes.

	LOCUS						
	Got-2	*α-Gpdh*	*Mdh*	*Adh*	*Dip-A*	*Hex-C*	*α-Amy*
Number of mobility mutants							
SM1, Cy/1(AW)	0/0	0/0	0/2	0/0	0/0	1/0	0/1
SM1, Cy/1(JH)	0/0	0/0	0/0	0/0	0/0	0/0	0/0
Number of null mutants							
SM1, Cy/1(AW)	0/0	1/1	0/0	0/0	0/0	0/0	0/0
SM1, Cy/1(JH)	0/1	4/1	0/0	0/0	0/0	4/0	0/0
Per locus mobility mutation rate ($\times 10^{-6}$)	0	0	4.63	0	0	2.31	1.88
Per locus null mutation rate ($\times 10^{-6}$)	2.31	16.49	0	0	0	9.26	0

Average number of allele generations per locus = 111,128
Total number of allele generations = 3,111,598
Average mobility mutation rate = $4/3{,}111{,}598 = 1.28 \times 10^{-6}$
Average null mutation rate = $12/3{,}111{,}598 = 3.86 \times 10^{-6}$
Total mutation rate = $16/3{,}111{,}598 = 5.14 \times 10^{-6}$

electrophoresis detects half of the amino acid variants, the amino acid replacement rate is about 11.2×10^{-6} per gene per generation, and since synonymous mutations are not seen, the rate including synonymous change is about 50% greater again. At the nucleotide level, the mutation rate is about 10^{-8} substitutions per nucleotide per generation. Because the haploid genome in man contains 3×10^{9} nucleotides, this implies that there are about 30 new nucleotide mutations per human gamete. Neel et al. (1988) report that the children of parents who were exposed to (and survived) the bomb's radiation (less than 2000 m from the hypocenter) had no higher mutation rate than children of individuals further than 2500 meters from the hypocenter: 3 out of 667,404 locus tests revealed mutations, yielding a rate of 4.5×10^{-6} per gene per generation. The reason there was no detectable increase in mutation rate among children of bomb survivors is that doses resulting in appreciable increases in mutation rate are lethal, and lower doses require enormous sample sizes to detect significant increases in mutation rates.

Is the Mutation Rate Adaptive?

There is evidently a wide range in rates of mutation. DNA repair mechanisms provide direct evidence of nature's attempts to regulate the rate of some

types of mutation. Later in this chapter we will see that mutations have a characteristic distribution of effects on fitness; many mutations are deleterious, but a lucky few are beneficial. Is it because of these lucky few that mutations occur, or is a baseline mutation rate inherent in the DNA replication and repair processes? Some organisms could evidently withstand a much higher mutation rate and still be able to compete. Why do they not have a higher mutation rate? Would an increased rate accelerate their rate of adaptive evolution? From the standpoint of population genetics, we must ask whether there is genetic variation in the rate of mutation, and whether individuals with high versus low rates differ in fitness. One experimental approach to this problem was to examine the dose-response relation for X-ray induced mutations in *Drosophila melanogaster* (Nöthel 1987). Populations that had been exposed to high doses of X-rays every generation for up to 600 generations had an X chromosome loss rate of about 6%. Surprisingly, a control population that did not receive X-rays had a loss rate of the X chromosome that was three times as high. Other results have confirmed a genetic response to chronic radiation dosage that lowered the rate of mutation per unit radiation dose.

Theoretical approaches to the problem of whether mutation rates are adaptive have shed some light on these important issues (Gillespie 1981a,b; Holsinger and Feldman 1983b; Holsinger et al. 1986; Liberman and Feldman 1986). The basic approach is to posit the existence of a modifier gene, different alleles of which result in different mutation rates of a target gene. The modifier itself is assumed to be neutral. Under these assumptions, with a constant array of fitnesses (constant environment), the lower mutation rate allele will always go to fixation. This result would seem to imply that the lower the mutation rate, the better, but there are theoretical conditions that favor an elevated mutation rate, including fluctuating environmental effects on fitness.

RATE OF CHANGE OF ALLELE FREQUENCY UNDER MUTATION PRESSURE

We now return to more familiar sorts of questions in population genetics and ask how the process of mutation affects the distribution and structure of genetic variation in populations. Consider a gene with two alleles A and a. Suppose that the mutation rate per generation from A to a is μ, and that the rate from a to A is ν. Let p_t and q_t (with $q_t = 1 - p_t$) denote the allele frequencies of A and a in generation t. Assuming that an allele cannot mutate twice in one generation, an A allele in generation t can originate in only two

ways: it could have been an A allele in generation $t - 1$ that escaped mutation to a (probability $1 - \mu$), or it could have been an a allele that mutated to A (probability ν). Symbolically,

$$p_t = p_{t-1}(1 - \mu) + (1 - p_{t-1})\nu \tag{3.1}$$

The solution of this equation gives p_t in terms of p_0 (the initial frequency of A), which is

$$p_t = \frac{\nu}{\mu + \nu} + \left(p_0 - \frac{\nu}{\mu + \nu}\right)(1 - \mu - \nu)^t \tag{3.2}$$

To understand the biological meaning of Equation 3.2, consider first the case when t is small — for example, fewer than 100 generations. Because μ and ν are ordinarily very small, the term $(1 - \mu - \nu)^t$ is approximately equal to $1 - t(\mu + \nu)$. Suppose also that $p_0 = 0$ (i.e., that the initial population is fixed for a). With these provisions (small t, $p_0 = 0$), Equation 3.2 implies that

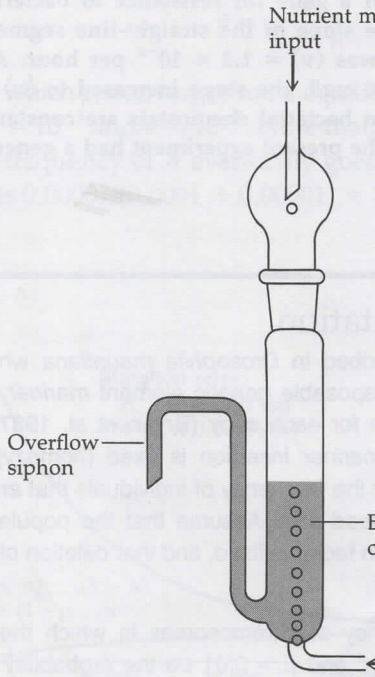


Figure 6. Diagram of a bacterial chemostat. Nutrient medium drips in at the top, but a constant volume is maintained by means of an overflow siphon. The air coming in at the bottom provides oxygen. Cells within the chemostat are in a continuous state of division, but the population does not increase in size because, in any given interval of time, the number of new cells produced by division is balanced by the number washed out through the siphon.

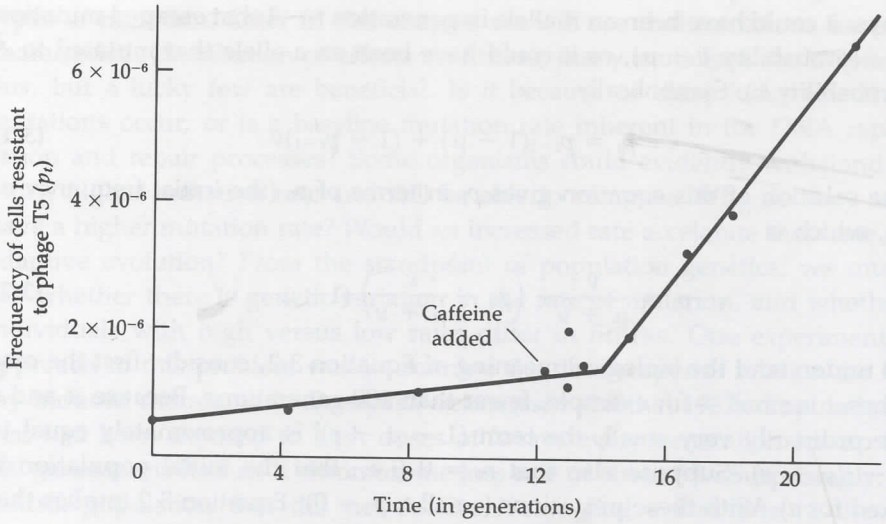


Figure 7. Measurement of mutation rates in bacterial chemostats. This example concerns the rate of mutation of a gene for resistance to bacteriophage T5. The mutation rate is estimated as the slope of the straight-line segments. Prior to the addition of caffeine, the slope was $\langle \nu \rangle = 1.3 \times 10^{-8}$ per hour. After addition of caffeine at a concentration of 150 mg/l, the slope increased to $\langle \nu \rangle = 12 \times 10^{-8}$ per hour. Mutation rates per hour in bacterial chemostats are constant for generation times between 2 and 12 hours. The present experiment had a generation time of 5.5 hours. (From Novick 1955.)

A The Dynamics of Mutation

A genetic factor has been described in *Drosophila mauritiana* which results in the spontaneous deletion of the transposable genetic element *mariner* at a frequency of approximately 1% per generation for each copy (Bryan et al. 1987). In a population containing a locus at which a *mariner* insertion is fixed (homozygous), how many generations would be required for the frequency of individuals that are homozygous for a deletion of the element to exceed 5%? Assume that the population is large, that mating is random, that the excision factor is fixed, and that deletion of the element does not affect survival or reproduction.

ANSWER Let p_t be the frequency of chromosomes in which the *mariner* element remains undeleted in generation t, and $\mu = 0.01$ be the probability of deletion of the element per generation. With $\nu = 0$ and $p_0 = 1$, Equation 3.2 implies that $p_t = (1 - \mu)^t$. The frequency of deletion homozygotes is greater than 5% when $(1 - p_t)^2 > 0.05$, or $p_t < 1 - (0.05)^{1/2} = 0.776$. Thus, t should be greater than $\ln(0.776)/\ln(0.99) = 25.2$ generations.

$p_t = t\nu$. That is, the frequency of the A allele increases linearly with time, and the slope of the line equals ν. Because ν is small, however, the linear increase in p_t is difficult to detect experimentally except in very large populations. In the laboratory, large population sizes can be attained in chemostats, which are devices for maintaining populations in a continuous state of logarithmic growth (Figure 6). The linear increase in p_t that occurs in bacterial chemostats from mutation pressure is shown in Figure 7. Note the abrupt increase in mutation rate (indicated by the increase in slope) shortly after the addition of caffeine, a bacterial mutagen.

To see what happens to allele frequency in the long run, consider Equation 3.2 when t is very large, for example 10^5 or 10^6 generations. Even though $1 - \mu - \nu$ is ordinarily close to 1, the value of t eventually becomes so large that $(1 - \mu - \nu)^t$ approaches zero. Thus, the whole right hand term in Equation 3.2 drops out, and p_t attains a value that remains the same generation after generation. Such a value of p is called an equilibrium value, which we will denote by $\hat{p}$. In the case of mutation, the equilibrium is given by

$$\hat{p} = \frac{\nu}{\mu + \nu} \tag{3.3}$$

The manner in which p_t converges to its equilibrium value is shown in Figure 8 for the case $\mu = 10^{-4}$ and $\nu = 10^{-5}$. Note that, whatever the initial frequency of A, the allele frequency of A eventually goes to $\hat{p}$, which in the example in the figure equals $0.00001/(0.0001 + 0.00001) = 1/11 = 0.091$. Figure 8 indicates

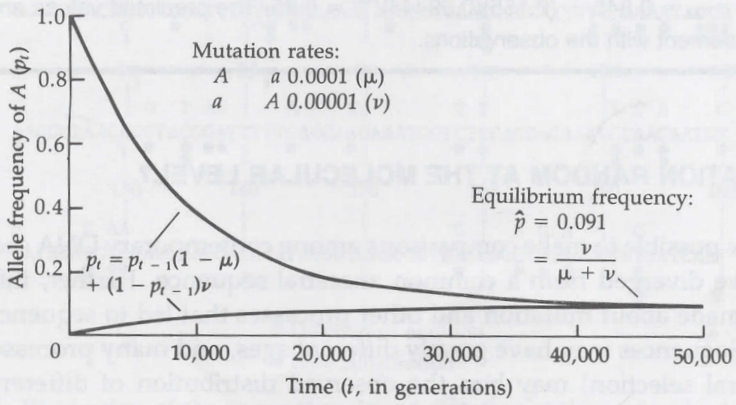


Figure 8. Theoretical change in allele frequency under pressure of reversible mutation. Note that attainment of near-equilibrium values requires tens of thousands of generations for realistic mutation rates.

that mutation pressure is usually very weak in changing allele frequency, inasmuch as the population requires thousands or tens of thousands of generations to reach equilibrium.

B Bidirectional Mutation

Stocker (1949) studied a case in the bacterium *Salmonella typhimurium* in which the mutation rates were sufficiently large that Equation 3.2 can be tested. The gene in question controls a protein component of the cellular flagella, and there are two alleles, which we can call *A* (for the "specific-phase" flagellar property) and *a* (for the "group-phase" flagellar property). The mutation rate from *A* to *a* was estimated as $\langle\mu\rangle = 8.6 \times 10^{-4}$ per generation, and that of *a* to *A* as $\langle\nu\rangle = 4.7 \times 10^{-3}$ per generation. These mutation rates are orders of magnitude larger than typically observed for other genes. The reason for these high mutation rates is that the change from *A* to *a* and back again does not involve mutation in the conventional sense, but results from intrachromosomal recombination (Simon et al. 1980). Formally, however, we can treat the system as one with reversible mutation. In cultures initially established with the frequency of *A* at $p_0 = 0$, Stocker found that the frequency increased to $p = 0.16$ after 30 generations and to $p = 0.85$ after 700 generations. In cultures initiated with $p_0 = 1$, the frequency decreased to 0.88 after 388 generations and to 0.86 after 700 generations. How do these values agree with those calculated from Equation 3.2 using the estimated mutation rates? What is the predicted equilibrium frequency of *A*?

ANSWER Note that $\langle\nu\rangle/(\langle\mu\rangle + \langle\nu\rangle) = 0.845$. This is the predicted equilibrium frequency (Equation 2.14). Also, $1 - \langle\mu\rangle - \langle\nu\rangle = 0.99444$, and this quantity determines the rate of approach to equilibrium. For the cultures with $p_0 = 0$, predicted values are $p_{30} = 0.845 - (0.845)(0.99444)^{30} = 0.13$ and $p_{700} = 0.845 - (0.845)(0.99444)^{700} = 0.83$. For the cultures with $p_0 = 1$, predicted values are $p_{388} = 0.845 + (0.155)(0.99444)^{388} = 0.86$ and $p_{700} = 0.845 + (0.155)(0.99444)^{700} = 0.85$. The predicted values are in very good agreement with the observations.

IS MUTATION RANDOM AT THE MOLECULAR LEVEL?

It is now possible to make comparisons among contemporary DNA sequences that have diverged from a common ancestral sequence. Further, inferences can be made about mutation and other processes that led to sequence divergence. Sequences may have greatly different ages, and many processes (such as natural selection) may bias the observed distribution of differences. To lessen the bias of selection, we will consider studies of nontranscribed genes, including heterochromatic sequences (Brown and Clegg 1983) and **pseudogenes** (genes that are no longer expressed but bear sequence similarity to active genes; Li et al. 1984).

Brown and Clegg (1983) examined sequences of a 176-bp segment of 12 copies of the knob heterochromatin of maize, and fitted models to the data to determine the distribution of the mutations (Figure 9). The first question one might ask is whether the distribution of mutational sites and their frequency is random. In this instance, because of the rareness of each mutation, a random distribution would be a Poisson distribution. The number of sites with 0, 1, 2, 3, and 4 mutations were 139, 24, 11, 1, and 1 respectively. The expected numbers under the assumption of a Poisson distribution with the same mean as that observed are 130.24, 39.21, 5.90, 0.60, and 0.04, and the χ^2 goodness-of-fit statistic is 12.87. With one degree of freedom (3 classes after pooling the smallest classes minus 1 minus 1 parameter estimated), this χ^2 has a very low probability ($P < 0.001$), so the fit to a Poisson is not good. The poor fit could also be seen by noting that the observed mean is 0.30 mutations per site, and the variance is 0.44, which is larger than the mean (the mean and variance are equal in a Poisson distribution). Brown and Clegg also performed run tests to determine whether the mutations were distributed evenly along the sequence, and they found that there was significant clustering of mutations.

The nonrandomness of mutations becomes especially pronounced when a substitution matrix among bases is examined (Table 4). Looking first at which bases mutated, the right column of the table shows that adenine was

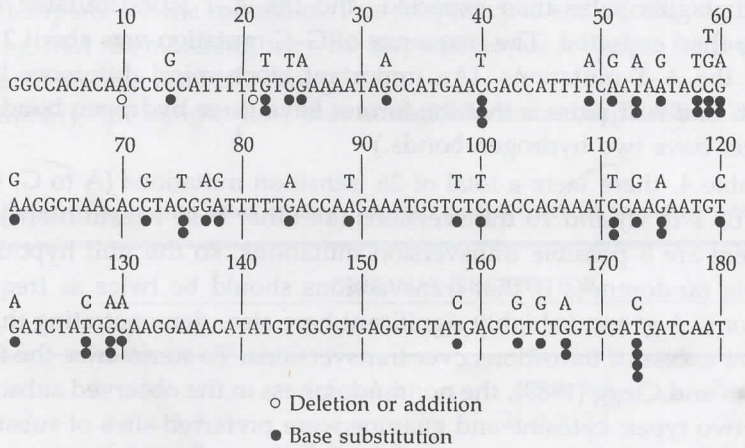


Figure 9. Illustration of the nonuniformity of mutation at the molecular level. The sequence is the consensus nucleotide sequence of maize satellite DNA associated with knob heterochromatin. Dots below the sequence indicate the position and frequency of variants, and the variant nucleotide at the site is indicated above the sequence. (From Brown and Clegg 1983.)

Table 4. Observed single-base differences from the consensus maize satellite sequence.

CONSENSUS SEQUENCE		BASE INTRODUCED				
Base replaced	Number	A	C	G	T	Total
A	58	—	0	5	1	6
C	38	1	—	5	8	14[a]
G	37	9	0	—	1	10
T	43	2	4	0	—	6
Totals	176	12	4	10	10	36

[a] C substituted twice at position 59.
(From Brown and Clegg 1983.)

replaced 6 times, cytosine 14 times, guanine 10 times, and thymine 6 times. The expected frequency of these mutational events is calculated from the base composition: if adenine is twice as prevalent in the sequence, then adenine should mutate twice as frequently under the null hypothesis that mutations are random and independent of the base. The goodness-of-fit of the observed and expected is again tested by a χ^2, and here the value of 9.56 leads to rejection of the null hypothesis, meaning that the mutations were not random. In this case, both guanine and cytosine were observed to mutate at much higher rates than expected, and the A–T pairs mutated less frequently than expected. The frequency of G–C mutation was about 2.7 times that of the A–T mutations. (An important biochemical difference between the G–C and A–T pairs is that the former have three hydrogen bonds, while the latter have two hydrogen bonds.)

In Table 4, there were a total of 26 transition mutations (A to G, G to A, C to T or T to C) and 10 transversions (all other base substitutions). Recall that there are 8 possible transversion mutations, so the null hypothesis of complete randomness is that transversions should be twice as frequent as transitions. A χ^2 test is highly significant here also, demonstrating that there is a great excess of transitions over transversions. To summarize the findings of Brown and Clegg (1983), the nonrandomness in the observed substitutions was of two types: cytosine and guanine were preferred sites of substitution, and transitions were significantly more frequent than transversions.

Table 5 shows a matrix of base substitution rates compiled by Li et al. (1984) from sequence data of pseudogenes in 16 mammals. Because pseudogenes are not transcriptionally active, it is believed that the DNA sequences

are free to diverge without any constraints; differences that are seen among pseudogenes may represent mutational differences with little bias. Table 5 shows that transitions, which comprise 59.3% of the mutations, are nearly twice as frequent as expected with complete randomness — a highly significant result. Li et al. also found that G–C pairs mutated more frequently than A–T pairs, and suggested that the methylation of cytosine in C–G dinucleotides may account for some of this bias. The studies of Brown and Clegg (1983) and of Li et al. (1984) are among many studies that have shown that nucleotide substitution is decidedly a nonrandom phenomenon. Further analysis of patterns of substitution led Wu and Maeda (1987) to conclude that the rate of mutation for sequences in the β-globin region differs between the sense and the anti-sense DNA strand.

Other factors also influence the degree of nonrandomness of mutations. For example, long stretches of A–T nucleotide pairs result in a more relaxed DNA helix, exposing the nucleotides to the damaging effects of chemicals to a greater extent. Neighboring bases also affect the mutation rate. Koch (1971) found that in phage T4, UAG mutated to UGG at a rate 23 times higher than the rate at which UAA mutated to UGA, despite the fact that both events are an A to G transition. Evidently the presence of the G in the third position greatly increased the mutation rate relative to having an A in the third position, perhaps due to local conformation of the DNA. Golding (1987) examined base substitutions among alleles of alcohol dehydrogenase, chorion proteins, gamma globin, and actin and found that about 7% of substitutions had a nearby repeat or nearby palindrome (reversed repeat) capable of serving as a template for the mutation. This frequency is statistically significant, suggesting that mismatching may be responsible for a higher frequency of mutations than was formerly thought.

In summary, the simple view that mutations occur completely at random

Table 5. Relative mutation rates among the four nucleotides in pseudogenes.[a]

Mutant nucleotide	ORIGINAL NUCLEOTIDE			
	A	T	C	G
A	—	4.4 ± 1.1	6.5 ± 1.1	20.7 ± 2.2
T	4.7 ± 1.3	—	21.0 ± 2.1	7.2 ± 1.1
C	5.0 ± 0.7	8.2 ± 1.3	—	5.3 ± 1.0
G	9.4 ± 1.3	3.3 ± 1.2	4.2 ± 0.5	—

[a] Values based on substitution data from 16 pseudogenes in mammals. (From Li et al. 1984.)

must be rejected. At the nucleotide level, many examples of nonrandom mutation events have been described. The most widely cited nonrandomness is the transition bias, which results in a much higher frequency of transition mutations than would be expected if all nucleotide substitutions were equally likely.

Molecular mechanisms of mutation tell us nothing about their phenotypic effects. From the point of view of adaptive evolution, it is crucial to know the distribution of mutational effects on the phenotype. This is the topic of the next section.

EFFECTS OF SPONTANEOUS MUTATIONS ON FITNESS

The experiments of Wallace (see Figure 5) show that balancer chromosomes can be used to make an entire chromosome homozygous, and the recessive effects of any gene on the chromosome will then be expressed. In the study by Wallace (1958), recessive lethal mutations were of interest. Terumi Mukai, in a series of large-scale experiments beginning in the early 1960s, used a similar technique but with a nice twist to it. He was interested in the distribution of effects on viability of newly arising mutations. In order to address this, he constructed 104 lines of *Drosophila melanogaster* that were isogenic for the same second chromosome. These were then crossed to *Cy*/*Pm* females each generation (Figure 10). Note that recessive mutations can accumulate on the wildtype second chromosomes, designated by +, since they are held in a heterozygous state over the *Pm* (*Plum* eye color) chromosome. In generations 10, 20, 25, 32, 52, and 60, the cross *Cy*/+ × *Cy*/+ was performed for each of the lines, and the offspring were counted. The proportions of zygotes from this cross are 1 *Cy*/*Cy* : 2 *Cy*/+ : 1 +/+, but since the *Cy* homozygotes die as embryos, the expected offspring flies are a ratio of 2 *Curly*-winged to 1 wildtype. As before, if a recessive lethal mutation had occurred on the + chromosome, only *Curly*-winged offspring would be recovered. The number of lines that only produce *Curly*-winged progeny from the test cross provides an estimate of the accumulated lethal mutations among the lines. More subtly, if a recessive mutation occurred on the + chromosome that made the fly less viable (but was not lethal), so that a greater fraction would die between the time of zygote formation and censused adult, the reduced viability would distort the expected 2 : 1 ratio to, say, 2 : 0.9, or 2 : 0.8. Taking the ratio of twice the number of wildtype to the number of curly winged flies, one gets an estimate of the relative viability of the wildtype to the balancer stock. [Actually, a minor adjustment gives the estimate better properties (Haldane 1956), which will be discussed in Chapter 4.] This estimate is valid for comparing across lines provided the mutations are recessive, so that $Cy/+_i$ and $Cy/+_j$ have the same viability.

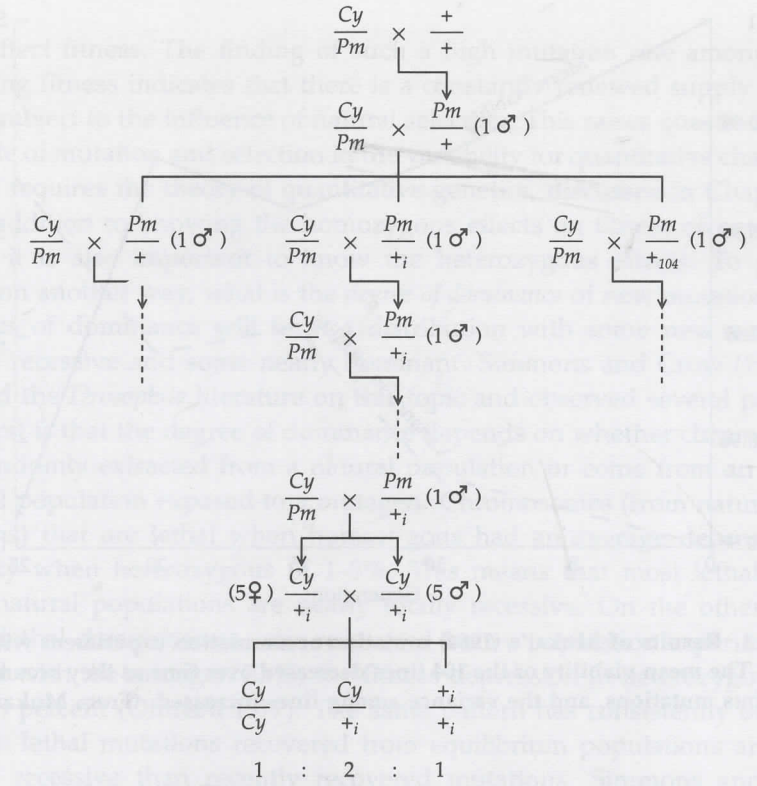


Figure 10. Mating scheme for accumulating mutations on the second chromosome of *Drosophila melanogaster*. The design is an extension of that in Figures 3 and 5, except that chromosomes are held in a heterozygous state for many generations, allowing mutations to accumulate. Periodically a test cross is done, as indicated at the bottom of the figure, in order to test the relative viability effects of the wildtype (+) chromosome. (From Mukai 1964.)

The results of Mukai's (1964) experiment appear in Figure 11. Several features are evident, including a decrease in the mean viability of the homozygous lines, and an increase in the variance among the lines. There were 15 lethal lines at generation 25, which works out to a mutation rate to lethals of about 0.006 per second chromosome per generation, a figure very close to that of Wallace (1958). Mukai was interested in estimating the rate of mutation of "polygenes," or genes with small effect on viability. To do this he used the following reasoning: assume that mutations are relatively rare events that occur with a Poisson distribution. Then the average number of mutations $\bar{p}$ is equal to the variance in the number of mutations. Letting $\bar{a}$ be the average effect on viability of a random mutation, we find that the expected rate of decrease of mean fitness across lines is $\bar{a}\bar{p}$. Similarly, the variance in the viability effects of the mutations equals $\bar{a}^2\bar{p}$. Over the first 25 generations of

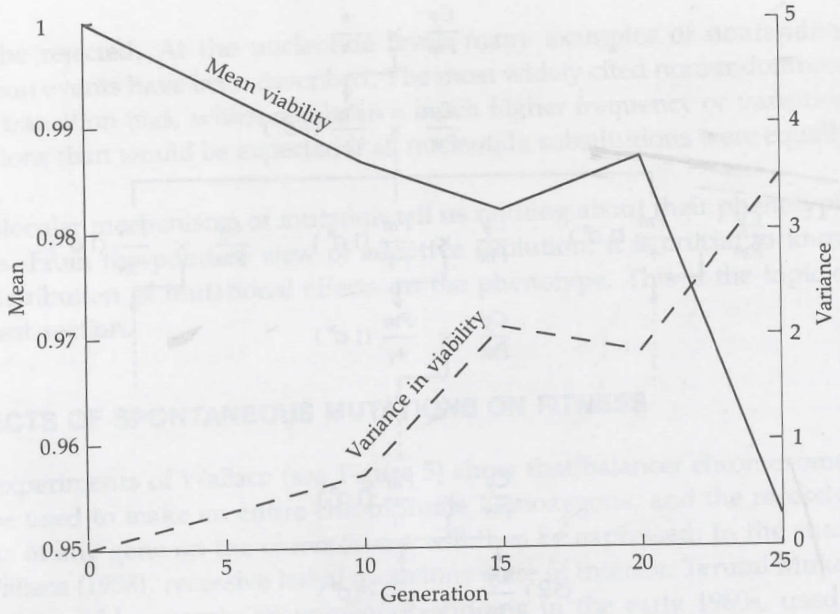


Figure 11. Results of Mukai's (1964) mutation accumulation experiment with *Drosophila*. The mean viability of the 104 lines decreased over time as they accumulated deleterious mutations, and the variance among lines increased. (From Mukai 1964.)

the Mukai experiment, the mean viability decreased almost linearly, and the variance increased almost linearly, consistent with the model that mutations of about equal additive effect accumulated at a constant rate. Regressions of the data in Figure 11 gave a slope of -0.1261 for the mean viability (per generation) and 0.1127 for the variance in viability. Since $\bar{a}\bar{p}$ is the expected rate of change of the mean, and $\bar{a}^2\bar{p}$ is the expected rate of change of the variance, we set

$$\bar{a}\bar{p} = 0.1261$$

$$\bar{a}^2\bar{p} = 0.1127$$

Solving for $\bar{a}$ and $\bar{p}$ we get $\bar{a} = 0.8937$ and $\bar{p} = 0.1411$. Hence the average number of new mutations that affect viability per second chromosome per generation is 0.1411 (about 1/7) and the average viability effect of each mutation is to reduce the viability to 0.8937, or 89% of the control viability. Note that the estimated mutation rate is about 30 times the mutation rate to lethal alleles.

Mukai's studies of spontaneous mutations were significant for a number of reasons. One important feature was that the mutations studied were ones

that affect fitness. The finding of such a high mutation rate among genes affecting fitness indicates that there is a constantly renewed supply of variation subject to the influence of natural selection. This raises questions about the role of mutation and selection in the variability for quantitative characters, which requires the theory of quantitative genetics, discussed in Chapter 8.

In addition to knowing the homozygous effects on fitness of new mutations, it is also important to know the heterozygous effects. To put the question another way, what is the *degree of dominance* of new mutations? The degrees of dominance will form a distribution with some new mutations totally recessive and some nearly dominant. Simmons and Crow (1977) reviewed the *Drosophila* literature on this topic and observed several patterns. The first is that the degree of dominance depends on whether chromosomes are randomly extracted from a natural population or come from an experimental population exposed to a mutagen. Chromosomes (from natural populations) that are lethal when homozygous had an average depression in viability when heterozygous of 1–3%. This means that most lethal alleles from natural populations are nearly totally recessive. On the other hand, when lethal chromosomes were recovered from a population experimentally treated with the mutagen EMS, the fitness depression in heterozygotes was 3.6–5.6 percent (Ohnishi 1977). The same pattern has consistently been observed: lethal mutations recovered from equilibrium populations are more nearly recessive than recently recovered mutations. Simmons and Crow (1977) interpret this pattern as evidence that natural selection rapidly eliminates lethals with deleterious effects in heterozygotes.

Some chromosomes, while not lethal, suffer a dramatic reduction in viability when made homozygous by the balancer chromosome method. These deleterious or subvital chromosomes can also be tested as heterozygotes to determine their degree of dominance. For these chromosomes, the reduction in heterozygous viability for deleterious alleles from equilibrium populations is 20–35% of the effect in homozygotes (i.e., if the homozygote viability is 0.8, the heterozygote viability is 0.96). However, when the deleterious chromosome is recovered from a mutagenesis experiment, the fitness reduction in heterozygotes is 35–50% of that in homozygotes (Simmons and Crow 1977). The conclusion is that mildly deleterious chromosomes have a much higher degree of dominance than lethals, and that new mutations have a higher degree of dominance than mutations drawn from populations in steady state. These effects balance in such a way that the absolute viabilities of heterozygotes for a lethal or a deleterious chromosome are about the same. Since most such mutations are rare, they occur primarily in heterozygotes, so lethals and deleterious alleles reduce the mean fitness of the population to about the same extent.

One final point about the effects of mutations on fitness is that fitness depends on more than just the progeny ratios observed in the experiments. The differences in offspring ratios reflect differences in the ability to survive from zygotes to adults. Hence, the only aspect of fitness assayed is viability. However, mutations may also cause differences in fertility components, such as mating ability or egg production. Sved and Ayala (1970), Sved (1971), and Tracey and Ayala (1974) examined viability as described above, but they also measured the effects of chromosome homozygotes on *total* fitness by allowing competition to occur among lines of *Drosophila pseudoobscura* in the laboratory. Some of the details of estimating fitness from changes in chromosome frequencies will be dealt with in Chapter 4, but for now the results can be summarized by noting that the average homozygote viability was 0.73, and the average net homozygote *fitness* was only 0.14. In other words, the deleterious effects estimated in viability tests represent only about one-third of the total deleterious effect.

INFINITE ALLELES MODEL

Balance Between Mutation and Random Genetic Drift

Recall from Chapter 1 that genes encoding enzymes often have more than two alleles. It is therefore of some importance to determine how many alleles can be maintained by mutation pressure. If genes tend to have more alleles than would be expected from mutation pressure alone, then other forces that operate in nature must tend to favor the accumulation of alleles. On the other hand, if there are fewer alleles than expected, then other forces must tend to eliminate alleles. There is a technical problem in calculating the number of alleles that can be maintained by mutation. If the procedure for distinguishing among alleles has low resolving power, few alleles will be detected no matter how many may actually be present in the population. Protein electrophoresis augmented with other procedures may have satisfactory resolving power for some types of studies, but many alleles may still remain undetected. Nevertheless, in principle, we can assume that all alleles are distinguishable, as is the case when their DNA nucleotide sequences are determined. Since an average protein contains roughly 300 amino acids, the average length of the amino acid coding part of a gene is roughly 900 nucleotides. With four possible nucleotides at each site, the number of possible alleles is about $4^{900} \approx 10^{542}$. Hence, we can suppose that every time a mutation occurs, it creates a new allele that does not already exist in the population. This is called the **infinite alleles model** of mutation. The infinite alleles model is but one way to specify the characteristics of new mutations.

Although the model represents a somewhat simplified view of mutation, it nevertheless provides a useful standard of comparison for other models or for observed allele frequencies.

In the infinite alleles model, we must also assume that the population in question is finite, because in an infinite population, new alleles would accumulate forever. Consequently, in the infinite alleles model both random genetic drift and mutation occur. In order to evaluate whether actual populations contain more or fewer alleles than would be expected from mutation pressure alone, we must determine how many alleles can be maintained in the infinite alleles model. That is, for what number of alleles is the creation of new alleles by mutation balanced exactly by the loss of old alleles due to random genetic drift? For the infinite alleles model, the answer is surprisingly simple. It can be calculated by expressing the proportion of homozygous genotypes in two ways: in terms of the allele frequencies, and in terms of the fixation index F. Setting these two proportions equal to each other yields the desired result. Accordingly, let the effective population number be N and the mutation rate to neutral alleles be μ. Because of the infinite alleles assumption, each allele in the population arises only once, and homozygotes for any allele are therefore autozygous. For n alleles with frequencies p_1, p_2, p_3, . . . , p_n, the homozygosity in any subpopulation is

$$\Sigma p_i^2 = p_1^2 + p_2^2 + \ldots + p_n^2 \qquad \textbf{(3.4)}$$

because mating is random within each subpopulation. While Σp_i^2 equals the homozygosity in terms of allele frequency, the autozygosity in the infinite alleles model also equals the fixation index F. The calculation of F_t that was derived in Chapter 2 (see Figure 8) is still correct for this model providing neither allele (α_i or α_j) undergoes mutation in the passage of one generation. Therefore,

$$F_t = [1/2N + (1 - 1/2N)F_{t-1}](1 - \mu)^2 \qquad \textbf{(3.5)}$$

The equilibrium value of F_t — call it $\hat{F}$ — is found by solving

$$\hat{F} = [1/2N + (1 - 1/2N)\hat{F}](1 - \mu)^2$$

from which

$$\hat{F} = \frac{1}{4N\mu + 1} \qquad \textbf{(3.6)}$$

to an excellent approximation (see Box C). Hence, the number of selectively neutral alleles increases under mutation pressure until $\hat{F}$ satisfies Equation 3.6.

C Equilibrium Autozygosity for the Infinite Alleles Model

From the assumptions of the infinite alleles model, we obtained the expression

$$F = [1/2N + (1 - 1/2N)F](1 - \mu)^2 \quad \textbf{(3.7)}$$

After some algebra, the exact solution $\hat{F}$ for the equilibrium autozygosity can be written

$$\hat{F} = \frac{(1 - \mu)^2}{2N - (2N - 1)(1 - \mu)^2} \quad \textbf{(3.8)}$$

If μ is very small (on the order of about 10^{-5}) then μ^2 is so small that we can ignore terms with this factor. Thus,

$$\hat{F} = \frac{1 - 2\mu}{4N\mu - 2\mu + 1} \quad \textbf{(3.9)}$$

For small μ, this is very close to

$$\hat{F} = 1/(4N\mu + 1) \quad \textbf{(3.10)}$$

If it is assumed that there is no mutation, but that migrants arrive from another population with the rate of arrival of new alleles from the other population of m per generation, then the equilibrium autozygosity can be written

$$F = [1/2N + (1 - 1/2N)F](1 - m)^2 \quad \textbf{(3.11)}$$

because $(1 - m)^2$ is the probability that neither of the two randomly chosen alleles comes from a migrant. It should be obvious, given the results from the infinite alleles model, that

$$\hat{F} = 1/(4Nm + 1) \quad \textbf{(3.12)}$$

When both migration and mutation are occurring, the equilibrium autozygosity is

$$F = 1/[4N(m + \mu) + 1] \quad \textbf{(3.13)}$$

We thus have two measures of homozygosity: Σp_i^2, which is the probability of drawing a pair of alleles identical by descent, and $\hat{F}$ in Equation 3.6. Since they measure the same thing, they must equal each other, so

$$\Sigma p_i^2 = \hat{F} = 1/(4N\mu + 1) \quad \textbf{(3.14)}$$

F is the probability of autozygosity, and H, which is often thought of as heterozygosity or virtual heterozygosity (Kimura 1983), is the probability of drawing a pair of alleles that are not autozygous. Since $H = 1 - F$, under the infinite alleles model, the equilibrium of H is

$$\hat{H} = \frac{4N\mu}{4N\mu + 1} \quad \textbf{(3.15)}$$

The relationship between the quantity $4N\mu$ and H is plotted in Figure 12. Note in particular that if $4N\mu = 1$, then $\hat{H} = 0.5$, and that higher mutation rates increase the equilibrium frequency of heterozygotes. (Alternative approaches leading to essentially the same result are discussed in Sved and Latter 1977.) One complication is that any number of distributions of allele frequency can result in the same homozygosity. For example, an equilibrium population with four alleles at frequencies $p_1 = 0.7$, $p_2 = 0.1$, $p_3 = 0.1$, $p_4 = 0.1$ has a homozygosity of $\hat{F} = p_1^2 + p_2^2 + p_3^2 + p_4^2 = (0.7)^2 + (0.1)^2 + (0.1)^2 + (0.1)^2 = 0.52$; likewise a population with two alleles at frequencies $p_1 = 0.6$, $p_2 = 0.4$ has a homozygosity of 0.52, because $p_1^2 + p_2^2 = (0.6)^2 + (0.4)^2 = 0.52$. The problem that many distributions of allele frequency can result in the same homozygosity can be sidestepped by assuming that all alleles are equally frequent. If the population contains n equally frequent alleles, then $p_1 = p_2 = p_3 = \ldots = p_n = 1/n$, and the homozygosity is given in Equation 2.15 as $n(1/n)^2 = 1/n$. At equilibrium, therefore,

$$1/n = \hat{F} = 1/(4N\mu + 1)$$

or

$$n = 4N\mu + 1$$

The number n of equally frequent alleles is called the **effective number of alleles**, often symbolized as n_e. Diverse distributions of allele frequency can therefore be compared in terms of their effective number of alleles. The four-allele population and the two-allele population that have identical homozy-

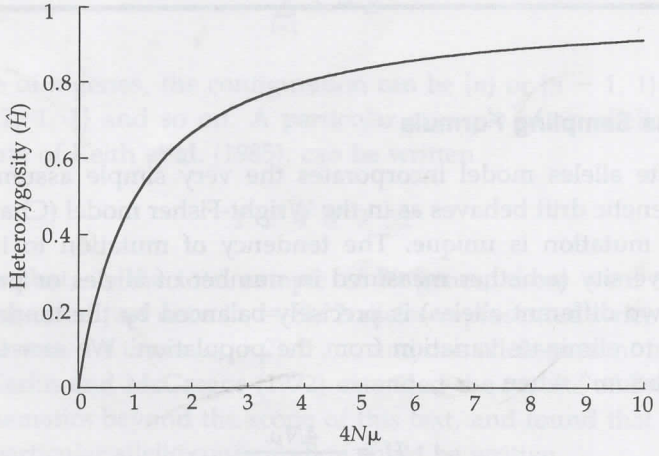


Figure 12. Prediction of steady-state heterozygosity under the infinite alleles model, $H = 4N\mu/(4N\mu + 1)$.

gosities of 0.52 have the same effective number of alleles, namely $n_e = 1/0.52 = 1.92$. Biologically speaking, n_e is the number of equally frequent alleles in an ideal population that would be required to produce the same homozygosity as in an actual population. At equilibrium, $n_e = 4N\mu + 1$.

D Effective Number of Alleles

In a large electrophoretic study of Caribbean populations of *Drosophila willistoni*, Ayala and Tracey (1974) found five alleles of the *leucine amino peptidase-5* gene (*Lap-5*), with allele frequencies estimated as 0.494, 0.429, 0.057, 0.014 and 0.006 in the Santo Domingo population, and 0.801, 0.177, 0.014, 0.004 and 0.004 in the Yunque population. The Yunque population contained eight alleles of the *xanthine dehydrogenase* gene (*Xdh*), with frequencies estimated as 0.446, 0.406, 0.092, 0.034, 0.014, 0.004, 0.002 and 0.002. The Yunque population had estimated allele frequencies of 0.574, 0.309, 0.114, and 0.003 of four alleles of the *adenylate kinase-1* gene (*Adk-1*).

a. Estimate the effective number of alleles in each of the above cases.
b. Show that an infinite population having a gene with an infinite number of alleles at frequencies 1/2, 1/4, 1/8, 1/16, 1/32, . . . has an effective number of alleles equal to 3. (Hint: The infinite sum $x + x^2 + x^3 + x^4 + \ldots$ equals $x/(1 - x)$ for any value of x less than 1.)

ANSWER **a.** For *Lap-5* in Santo Domingo, $\langle n_e \rangle = 2.32$; for *Lap-5* in Yunque, $\langle n_e \rangle = 1.49$; for *Xdh*, $\langle n_e \rangle = 2.68$; for *Adk-1*, $\langle n_e \rangle = 2.28$.
b. The homozygosity is $(1/2)^2 + (1/4)^2 + (1/8)^2 + \ldots$, which is $1/4 + 1/16 + 1/64 + \ldots$ or $x + x^2 + x^3 + \ldots$ with $x = 1/4$. Thus, the homozygosity equals $x/(1 - x) = (1/4)/(3/4) = 1/3$. The effective number of alleles is the reciprocal of the homozygosity, so $n_e = 3$.

The Ewens Sampling Formula

The infinite alleles model incorporates the very simple assumptions that random genetic drift behaves as in the Wright-Fisher model (Chapter 2), and that each mutation is unique. The tendency of mutation to increase the genetic diversity (whether measured in number of alleles or probability of drawing two different alleles) is precisely balanced by the tendency of random drift to eliminate variation from the population. We saw that there is an "equilibrium" when

$$H = \frac{4N\mu}{4N\mu + 1} \qquad (3.16)$$

This is not an equilibrium in the usual sense. In reality allele frequencies are always changing; new mutations continue to come into the population, and eventually they are always eliminated. The term *steady state* is probably more appropriate for this kind of behavior, since the actual alleles are not fixed at a constant frequency, but rather are entering and leaving the population. The population remains at a steady state in the sense that the number of alleles and the level of autozygosity remain fairly constant. If the number of alleles and the level of autozygosity do not change much, then it is reasonable to assume that there is also a steady-state distribution of allele frequencies. By this we mean that the most common allele always has a frequency of p_1, and the next most common has a frequency of p_2, and so on. The steady-state distribution is still stranger: even though the most common allele is expected to have a frequency of p_1, the *identity* of the most common allele is expected to change with time. In the steady-state population, not all alleles are equally frequent, and F is greater than it would be were all alleles equally frequent.

Consider the steady-state distribution of allele frequencies from the point of view of an experimenter taking a sample from a population. Let the sample size be n genes, and suppose there are k different alleles in this sample. The sample might consist of, for example, 10 unique alleles, 3 alleles that are represented twice in the sample, 7 alleles that are present 3 times, and so on. Such a description of the sample is called the allelic **partition** or **configuration**. We define a_i as the number of alleles that are present exactly i times in the sample. In a sample of size n, $0 < i \leq n$. Using this definition, note that

$$\sum_{i=1}^{n} a_i = k \tag{3.17}$$

In a sample of n genes, the configuration can be $\{n\}$ or $\{n - 1, 1\}$ or $\{n - 2, 2\}$ or $\{n - 2, 1, 1\}$ and so on. A particular example of an allelic partition, from the data of Keith et al. (1985), can be written

$$1^8\ 2^2\ 4^2\ 8^1\ 9^1\ 52^1$$

This means that, in the total sample of 89 genes, there were 8 singleton alleles represented just once ($a_1 = 8$), 2 alleles represented twice ($a_2 = 2$), 2 alleles represented 4 times ($a_4 = 2$), . . . , and one allele present in 52 copies ($a_{52} = 1$). Karlin and McGregor (1972) extended the results of Ewens (1972) using mathematics beyond the scope of this text, and found that the probability of a particular allelic configuration could be written

$$\Pr(a_1, a_2, \ldots, a_n) = \frac{n!\theta^k}{\theta(\theta + 1) \ldots (\theta + n - 1)} \prod_{i=1}^{n} \left(\frac{1}{i^{a_i} a_i!} \right) \qquad \textbf{(3.18)}$$

where $\theta = 4N\mu$. The expected configuration is entirely determined by the sample size, n and the number of observed alleles k. After some manipulation, the distribution of k, the number of alleles in the sample, is

$$\Pr(k \text{ alleles in sample of size } n) = |S_n^k| \theta^k / |S_n(\theta)| \qquad \textbf{(3.19)}$$

where $S_n(\theta) = \theta(\theta + 1)(\theta + 2) \ldots (\theta + n - 1)$, and $|S_n^k|$ is called a Stirling number of the first kind, which equals the coefficient of θ^k in $S_n(\theta)$ (Abramowitz and Stegun 1965, p.824).

The expectation of Equation 3.19 is the expected number of alleles in the sample, given θ and the sample size. The expectation is

$$E(k) = 1 + \theta/(\theta + 1) + \theta/(\theta + 2) + \ldots + \theta/(\theta + n - 1) \qquad \textbf{(3.20)}$$

If θ is very small, $E(k) \approx 1$, while for very large θ, $E(k)$ approaches n, implying that for a large enough population with a high enough mutation rate, every allele that we sample will be different. The form of Equation 3.20 suggests that as the sample size increases, more alleles will be found, but that there is a diminishing return in finding new alleles when the sample size gets large. When $E(k)$ is plotted against θ (Figure 13), the increase in the expected number of alleles is greatest for larger sample sizes. Karlin and McGregor (1972) also showed that if k alleles were found in a sample of n genes, then the probability that the next gene sampled will be a new allele is $\theta/(\theta + n)$.

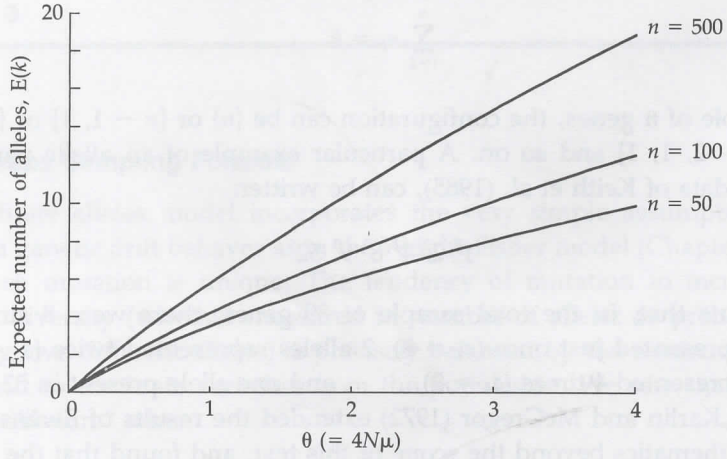


Figure 13. Relations between θ (= $4N\mu$), the expected number of alleles, and the sample size according to the Ewens-Watterson sampling theory of a population in steady state under the infinite alleles model of neutral mutation.

Our intuition is quite accurate here: as n increases, the chance of finding a new allele decreases, and in large populations with a high mutation rate, the probability that the next allele sampled will be new remains high for a larger sample size.

We already saw that the expected probability of autozygosity under the infinite alleles model is

$$F = 1/(\theta + 1) \qquad (3.21)$$

The infinite alleles model gives a steady-state prediction of F from θ as in Equation 3.21, and a prediction of k from Equation 3.20. Combining these, the expected relation between F and k is plotted in Figure 14. The hyperbolic relation is no surprise, because a population with many alleles will generally have a lower probability of identity of a randomly chosen pair of alleles. For $\theta = 1$ the expected F is 1/2 for all sample sizes, but the larger sample size should yield more distinct alleles.

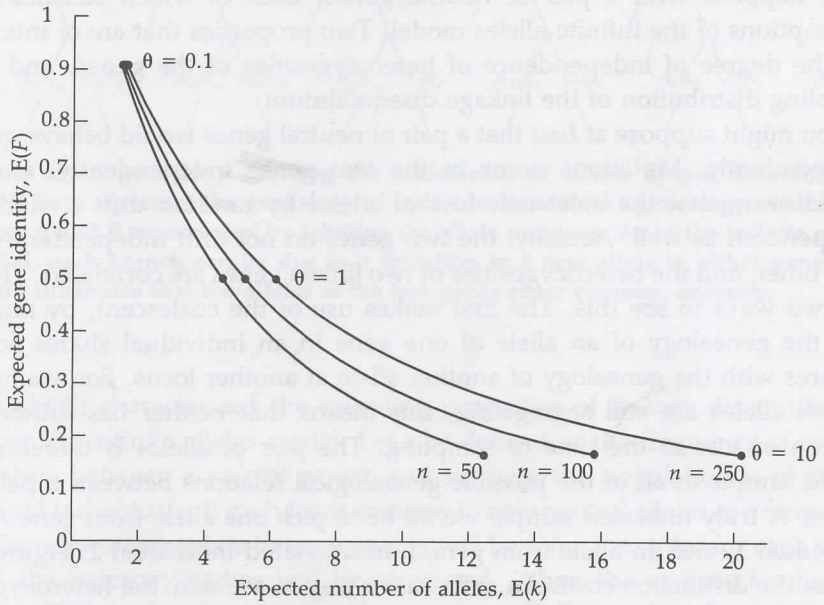


Figure 14. The infinite alleles model predicts this relation between the expected number of alleles and the expected gene identity F. The three curves represent a range of values of $\theta = 4N\mu$, starting at $\theta = 0.1$ in the upper left, and ending with $\theta = 10$ in the lower right. For the value of $\theta = 1$, the expected F, given by the relation $F = 1/(1 + \theta)$, is 1/2, regardless of the sample size. Larger sample sizes always lead to larger expected numbers of alleles, but the difference is greater in more diverse populations (those with smaller F).

Another formula of some interest, also due to Ewens, gives the expected configuration of k alleles in a sample of size n:

$$\mathrm{E}(a_i|k \text{ alleles in sample size } n) = \frac{n!|S_{n-i}^{k-1}|}{i(n-i)!|S_n^k|} \qquad (3.22)$$

This formula contains quite a surprise. Neither the population size nor the mutation rate needs to be known to test the fit between the observed configuration and the expected configuration given the infinite alleles model. This observation gave rise to a statistical test known as the **Ewens-Watterson test**, discussed further in the final section of this chapter.

Two Neutral Genes and the Infinite Alleles Model

The infinite alleles model provides a number of useful properties about the joint effects of mutation and random drift acting on a single gene. Real organisms have more than one gene, and just as we needed to consider more than one gene in the Hardy-Weinberg principle, it is instructive to see what happens with a pair of neutral genes, each of which satisfies the assumptions of the infinite alleles model. Two properties that are of interest are the degree of independence of heterozygosities of the genes, and the sampling distribution of the linkage disequilibrium.

You might suppose at first that a pair of neutral genes would behave quite independently. Mutations occur in the two genes independently, and it would seem that the stochastic loss of alleles by random drift should be independent as well. Actually, the two genes do not drift independently of each other, and the heterozygosities of two linked genes are correlated. There are two ways to see this. The first makes use of the coalescent, by noting that the genealogy of an allele of one gene in an individual shares some features with the genealogy of another allele at another locus. For example, if both alleles are still segregating, this means that neither has drifted to fixation or loss at the time of sampling. The pair of alleles is therefore a biased sample of all of the possible genealogical relations between a pair of alleles. A truly unbiased sample would be to pick one allele from gene A of individual 1, and an allele from gene B of unrelated individual 2. Figure 15 makes the distinction clear. An even easier way to see why the heterozygosities of neutral genes become correlated is presented in Box E.

If the heterozygosities are correlated, it stands to reason that the joint probabilities of pairs of alleles on chromosomes must not be independent. This is precisely what is meant by linkage disequilibrium. The infinite alleles model generates statistical associations between linked alleles, and the variance in the predicted distribution of D depends on the population size and the frequency of recombination. In an extensive series of simulations, Hud-

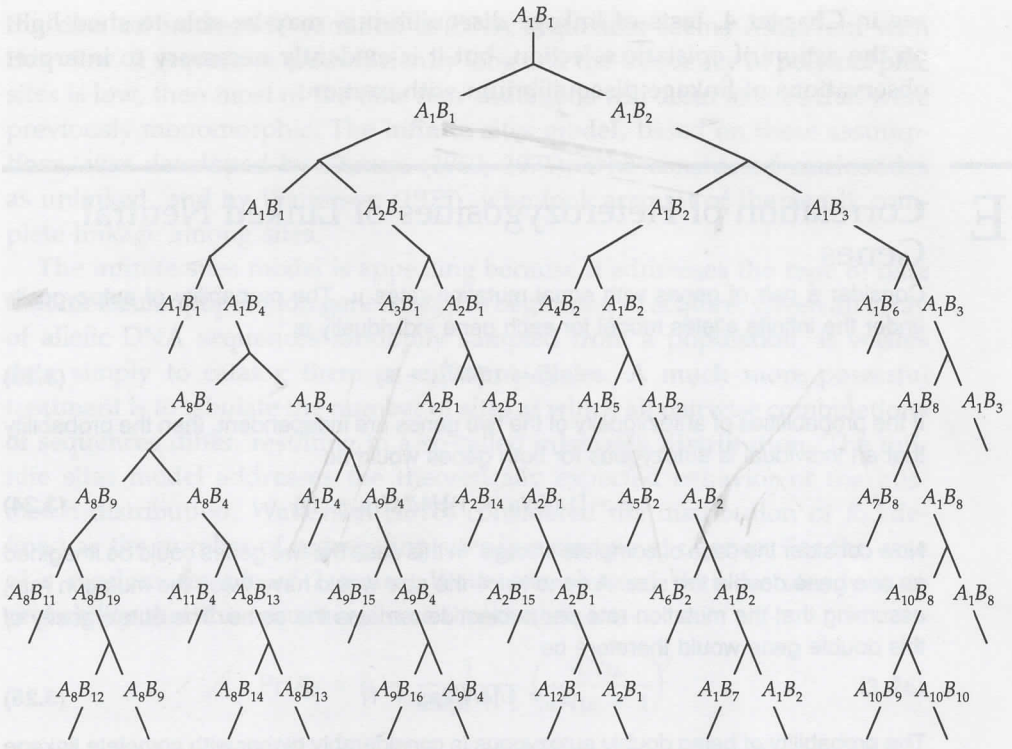


Figure 15. A genealogy showing the coalescence of alleles at two different loci. Black lines represent the genealogy of individuals, with the genealogy of alleles at genes *A* and *B* represented by labeling the allele numbers. As in the infinite alleles model, each branch can be due to a mutation to a new allele in either gene. The figure illustrates that the alleles at the two genes share common ancestry.

son (1985) characterized the sampling properties of linkage disequilibrium under the infinite alleles model. Let r be defined as the frequency of recombination between a pair of genes, and suppose the population is of size N diploid individuals. If each locus mutates to new neutral alleles at a frequency μ, then the steady-state heterozygosity will be $4N\mu/(1 + 4N\mu)$ for each gene, and the heterozygosities will be correlated. When the statistic for linkage disequilibrium is corrected for the number of alleles of each gene in the sample, then linkage disequilibrium is almost independent of $N\mu$. The amount of linkage disequilibrium does depend in a complex way on $4Nr$, which represents a combination of sampling and recombination. If $4Nr < 10$, then random drift can generate substantial amounts of linkage disequilibrium, and the variance in the estimate is so high that it is difficult to make inferences about other forces contributing to the disequilibrium. As we will

see in Chapter 4, tests of linkage disequilibrium may be able to shed light on the action of epistatic selection, but it is evidently necessary to interpret observations of linkage disequilibrium with caution.

E Correlation of Heterozygosities of Linked Neutral Genes

Consider a pair of genes with equal mutation rates μ. The probability of autozygosity under the infinite alleles model for each gene individually is

$$F = 1/(4N\mu + 1) \tag{3.23}$$

If the probabilities of autozygosity of the two genes are independent, then the probability that an individual is autozygous for both genes would be

$$F = [1/(4N\mu + 1)]\,[1/(4N\mu + 1)] \tag{3.24}$$

Now consider the case of complete linkage. In this case the two genes could be imagined as one gene double the size. A gene twice the size would have twice the mutation rate, assuming that the mutation rate per nucleotide remains the same. The autozygosity of this double gene would therefore be

$$F = 1/[4N(2\mu) + 1] \tag{3.25}$$

The probability of being doubly autozygous is considerably higher with complete linkage than if the autozygosities were independent. Given that the first gene is autozygous, the probability that the second one is also autozygous is $(1 + 4N\mu)/(1 + 8N\mu)$, which is greater than $1/(4N\mu + 1)$. Similarly, given that the first gene is heterozygous, the chance of heterozygosity of the second gene is $8N\mu/(8N\mu + 1)$. This result is equivalent to saying that the heterozygosities H for the two linked genes are correlated under the infinite alleles model. As $N\mu$ (the absolute number of new mutations per generation) decreases, and as the amount of recombination between the genes increases, the magnitude of the correlation decreases. Tajima (1983) showed that the correlation in heterozygosity of linked genes is equivalent to a correlation in the average number of nucleotide differences among alleles of the two genes, relating the problem to the infinite sites model discussed in the next section.

INFINITE SITES MODEL

Rather than considering each mutation as generating a unique allele, with an infinite number of possible alleles, we can instead consider an allele as a sequence of nucleotides with mutation altering a site in the sequence. If the mutation rate is sufficiently low, then most sites will be monomorphic, and all polymorphic sites will be segregating for just two nucleotides. Much of

the data on nucleotide variation in DNA sequences seems consistent with this view. If sequences are sufficiently long and the frequency of polymorphic sites is low, then most of the time new mutations will occur at sites that were previously monomorphic. The **infinite sites model**, based on these assumptions, was developed by Kimura (1969, 1971), who considered nucleotides as unlinked, and by Watterson (1975), who took account of the nearly complete linkage among sites.

The infinite sites model is appealing because it addresses the type of data that molecular population geneticists are beginning to acquire. Given an array of allelic DNA sequences randomly sampled from a population, it wastes data simply to catalog them as different alleles. A much more powerful treatment is to tabulate the number of sites at which all pairwise combinations of sequences differ, resulting in a so-called **mismatch distribution**. The infinite sites model addresses the theoretically expected behavior of the mismatch distribution. Watterson (1975) considered the distribution of K_i, defined as the number of segregating sites in a sample of i genes. For the case of a random sample of 2 genes, Watterson showed that the steady-state probability that the sequences have i mismatches is

$$\Pr(K_2 = i) = \frac{1}{4N\mu + 1}\left(\frac{4N\mu}{4N\mu + 1}\right)^i \qquad \textbf{(3.26)}$$

where μ is the mutation rate *per gene* (not per site). A particular case of this equation gives the probability that two sequences have no sites different, and hence are identical. This is

$$\Pr(K_2 = 0) = 1/(4N\mu + 1) \qquad \textbf{(3.27)}$$

in agreement with the infinite alleles model. The mean and variance in the distribution of number of segregating sites are $4N\mu$ and $4N\mu + (4N\mu)^2$ respectively. Note that the variance is larger than the mean because of the correlation among sites caused by linkage (see the previous section).

Kimura showed that the number of segregating sites expected under the infinite sites model is approximately $4N\mu \ln(2N)$. For example, in a population with 500,000 individuals ($2N = 10^6$) and a mutation rate of 10^{-6}, the expected number of segregating sites is 27.6. In any individual, the expected number of segregating sites (heterozygous sites) is $4N\mu$, in this case, 2. Since the number of alleles in a sample must depend on the sample size as well as the population size and mutation rate, the expected number of segregating sites in a sample of n genes must be smaller than $4N\mu \ln(2N)$. Watterson (1975) obtained an expression that accounts for this sampling effect.

Within a single gene over the span of one generation, mutation is a rare phenomenon, and intragenic recombination might similarly be rare. Since recombination shuffles the configuration of sites on chromosomes, it might

profoundly affect the expected behavior of the infinite sites model. We saw that some aspects of the infinite sites model parallel those of the infinite alleles model, and it would be useful to consider results for the expected steady state distribution of allele configurations under the infinite sites model with recombination. A two-site model with recombination was studied by Strobeck and Morgan (1978), who found by simulation that when $4N\mu > 1$, the variance in F exceeded the expected value from the infinite alleles model. The infinite alleles model therefore does not apply when there is intragenic recombination.

The Coalescent under the Infinite Sites Model

Under the infinite sites model, while most new mutations rapidly disappear from the population, a minority achieve appreciable frequency. For any population sampled at a particular time, the genealogy of an allele in the sample can be traced back to a single common ancestor. For any pair of individuals in the population, their genealogy can always be traced back to a common ancestor. The chance that two genes were identical by descent from a common ancestor in the preceding generation is $1/2N$, and the probability that they are not identical by descent is $1 - (1/2N)$ (assuming a diploid population with effective size N). The chance that the common ancestor occurred t generations back is obtained by multiplying $1/2N$ (the chance of having the common ancestor exactly t generations ago) by $(1 - 1/2N)^t$, (the chance of not having a common ancestor since). The probability distribution of time to the common ancestor is therefore the exponential

$$\Pr(\text{2 alleles had common ancestor } t \text{ generations ago}) =$$

$$(1/2N)[1 - (1/2N)]^t \approx (1/2N)\, e^{-t/(2N)} \qquad \textbf{(3.28)}$$

This distribution has a mean of $2N$ generations and a variance of $(4N)^2$. Note that the confidence interval around the mean time is not very tight, since the standard deviation of the distribution is equal to the mean.

In a sample of n genes, we can calculate the time to the most recent splitting of alleles. At time 0 there are n different lineages, and one generation ago there might also be n lineages. As we go further back in time, eventually we come to a fork where some pair of the alleles in the sample split from a common ancestor. The probability that this branch occurred exactly t generations ago is

$$\Pr(\text{first branch in genealogy of } n \text{ alleles occurred } t \text{ generations ago}) =$$

$$\frac{\binom{n}{2}}{2N} \exp[-\binom{n}{2} t/(2N)] \qquad \textbf{(3.29)}$$

where $\binom{n}{2} = n!/[(n-2)!2!]$ = the number of ways to get combinations of 2 items out of n. This distribution has a mean of $4N/[n(n-1)]$ generations and a variance of $16N^2/[n(n-1)]^2$. Again, the standard deviation is equal to the mean (Kingman 1982; Tajima 1983). Along the same lines, the expected time to the coalescence of all of the n alleles (i.e., the most recent time that one sample of n alleles shared a common ancestor) is

$$E(t) = 4N(1 - 1/n) \tag{3.30}$$

with variance

$$\text{Var}(t) = 4N^2 \sum_{i=2}^{n} [1/\binom{i}{2}^2]$$

(Kingman 1982; Tajima 1983). As the sample size n increases toward the total population size, t approaches $4N$, which equals the expected fixation time for a newly arisen mutation.

In the infinite sites model, any pair of alleles can be distinguished by the number of sites at which they differ. If S is the number of distinct sites distinguishing a pair of alleles drawn at random from the population, and T is the sum of the branch lengths of the tree back to coalescence, then S is geometrically distributed at steady state with

$$E(S) = 2\mu\, E(T) = 4N\mu \tag{3.31}$$

This suggests a means of testing the goodness-of-fit of a sample of alleles to the infinite sites model. If the mismatch distribution is constructed from the sequences of the alleles in the sample, the infinite sites model makes a well-defined prediction of the form of the mismatch distribution. Application of these ideas will be pursued further in Chapter 7.

CHARGE STATE MODEL

Motivated by the observation of distinct bands on electrophoretic gels caused by charge differences in proteins, Ohta and Kimura (1973) proposed a model in which mutations allow alleles to change by steps j ($j = 0, \pm 1, \pm 2, \ldots$) with probability μ_j. Since mutations are assumed to occur in discrete steps, this model is known as the **charge state** or **stepwise mutation model**. Suppose that in a population of size N there are N_i proteins of charge i ($i = 0, \pm 1, \pm 2, \ldots$). The sum of the N_i is $2N$. The probability π_i that a protein taken at random is of charge i is

$$\pi_i = \frac{\ldots + \mu_2 N_{i-2} + \mu_1 N_{i-1} + \mu_0 N_i + \mu_{-1} N_{i+1} + \ldots}{2N} \tag{3.32}$$

The model assumes a multinomial sampling of these alleles, and after a sufficiently long period of time, it attains a steady state analogous to that for the infinite alleles model. The case of $\mu_i = 0$ for $j \geq 2$ was considered by Ohta and Kimura (1973), restricting the changes to just one step in charge. For this case, they found that the expected probability of autozygosity was

$$E(F) = 1/(1 + 8N\mu)^{1/2} \qquad (3.33)$$

This gives a somewhat higher value for the probability of autozygosity than the infinite alleles model, but for small $N\mu$ the difference is not great. Attempts to fit actual electrophoretic data to the charge state model have not been very successful (Weir et al. 1976; Wehrhahn 1975; Ramshaw et al. 1979; Fuerst and Ferrell 1980), in part because charge is not the only property that affects electrophoretic mobility.

TESTS OF SELECTIVE NEUTRALITY

The hypothesis that observed allozyme polymorphisms result from selectively neutral alleles maintained in a balance between mutation and random genetic drift is known as the **neutrality hypothesis** or the hypothesis of selective neutrality (Kimura 1968; King and Jukes 1969). In essence, the neutrality hypothesis states that many mutations have negligible effects on the ability of organisms to survive and reproduce. The frequencies of neutral alleles are not influenced by the forces of natural selection, but are instead determined by other forces such as migration (discussed later) and random genetic drift. Because the neutrality hypothesis is of fundamental importance in population genetics and evolution, it has been a subject of considerable debate. If the hypothesis is generally true, then most observed polymorphisms — of allozyme genes, for example — may have no particular significance in the adaptation of organisms to their present environments. From this perspective, selectively neutral polymorphisms are mere evolutionary "noise," and regardless of how much their study may reveal about population structure and random genetic drift, they tell us little or nothing about adaptive genetic changes in evolution. If, on the other hand, the neutrality hypothesis is false, the next important task would be to study in detail the actual adaptive significance of genetic polymorphisms.

To assess the plausibility of the neutrality hypothesis, many aspects of the hypothesis must be compared with data collected from actual populations. One aspect of the hypothesis developed in the preceding section concerns the homozygosity to be expected with the infinite alleles model. Using observed homozygosities, we can estimate the effective number of alleles $\langle n_e \rangle$, and, from the expression $n_e = 4N\mu + 1$, estimate the corresponding values

of $\langle N\mu \rangle$. If the resulting values are grossly unreasonable, we can safely reject the infinite alleles version of the neutrality hypothesis (or at least argue that actual populations cannot be in equilibrium).

Chapter 1 summarized data showing that observed values of heterozygosity of allozyme genes range from 0.04 to 0.14 in most organisms. Observed homozygosities therefore range from $1 - 0.04 = 0.96$ to $1 - 0.14 = 0.86$, which corresponds to estimated $\langle n_e \rangle$ in the range $1/0.96 = 1.04$ to $1/0.86 = 1.16$. Estimates of $\langle N\mu \rangle$, calculated as $(\langle n_e \rangle - 1)/4$, therefore range from 0.01 to 0.04. That the maximum estimated value of $\langle N\mu \rangle$ differs from the minimum by a factor of only about 4 is surprising inasmuch as the population number in different species ranges over a factor of 10^4 or more. The distribution of homozygosities for allozyme genes thus appears to be too uniform among diverse organisms, and has been interpreted as meaning that the neutrality hypothesis is wrong. This conclusion implies that natural selection is involved somehow in the maintenance of genetic polymorphisms. On the other hand, rejection of the neutrality hypothesis on grounds of the estimated range of $\langle n_e \rangle$ is probably premature because routine electrophoresis does not distinguish all alleles. Beyond that, estimates of effective population number $\langle N \rangle$ in natural populations are generally imprecise because the studies are very difficult to conduct. Further, estimates of $\langle \mu \rangle$ (which in this case is the mutation rate to neutral alleles) are even more uncertain than $\langle n_e \rangle$.

Figure 16 shows a second type of test of the adequacy of the neutrality hypothesis in explaining observed levels of genetic variation of allozyme genes. The figure shows the observed distribution of allele frequencies of samples of genes in four different organisms (open columns) along with the computer-generated theoretical distribution expected with the infinite alleles model (black columns). The correspondence between the histograms is fairly good, but in the Japanese macaque (Figure 16a), there appears to be an excess of rare alleles. The correlation between the number of alleles and the sample heterozygosity was also consistent with the neutral expectation (Chakraborty et al. 1980).

A third test of the neutrality hypothesis is shown in Figure 17, which presents data on the mean and variance of heterozygosity in 77 vertebrate species. The curve is the theoretical expectation from the infinite alleles model when the rate of selectively neutral mutation varies among genes (Nei et al. 1976). The fit in Figure 17 is impressive, especially considering that the curves are not fits to the data but are strictly theoretical predictions. On the other hand, the observed points are sufficiently scattered that any number of other curves might fit at least as well. Evidently, statistical comparisons of this sort do not have sufficient power to firmly support or refute the neutrality hypothesis.

A brief background on the phrase *lacking in power* may be in order. The

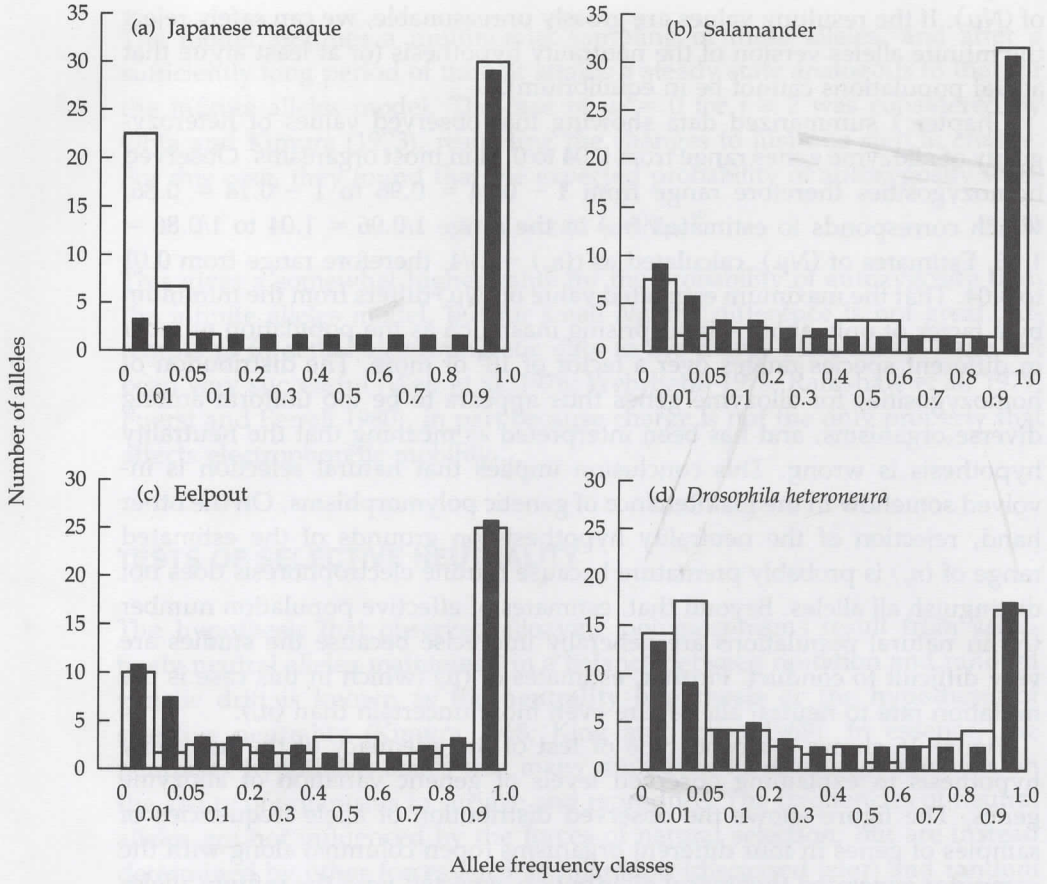


Figure 16. A statistical test of the neutral mutation theory. Observed and expected distributions of allele frequencies are plotted in four species representing different average heterozygosity values. (a) Japanese macaque (20 loci, average heterozygosity $\langle H \rangle = 0.018$, $n = 1976$). (b) *Taricha rivularis* (salamanders) (37 loci, $\langle H \rangle = 0.077$, $n = 784$). (c) *Zoarces viviparus* (eelpouts) (32 loci, $\langle H \rangle = 0.102$, $n = 757$). (d) *Drosophila heteroneura* (25 loci, $\langle H \rangle = 0.162$, $n = 605$). The observed distributions are represented by open columns, and the expected distributions by black columns. (From Chakraborty et al. 1980.)

neutral theory is useful as a starting point, or null hypothesis, which provides predictions about the relations among observed quantities that can be confirmed or rejected. Statistical tests of the neutral theory are similar to other types of statistical tests in that two distinct types of possible errors must be balanced. If the tests are too demanding, for example in failing to allow for the effects of random sampling error, then data may often result in rejection of the hypothesis, even when it is true. This is called Type I error. On the other hand, if the statistical test allows too much latitude in the data, then

data will seldom result in rejection of the hypothesis, even when it is false. This is called Type II error. The trade-off between Type I and Type II error is that the probability of Type I error cannot be decreased without increasing the probability of Type II error, and vice versa. By convention, statisticians usually adopt a 5% criterion for rejection of the null hypothesis. This is the familiar 5% level of statistical significance, and it means that there is a 5% chance of rejecting a true hypothesis (Type I error). With this convention, the probability of a Type II error (failing to reject a false hypothesis) falls where it may, and tests with relatively high probabilities of Type II error are said to be lacking in power.

Another test of the neutral theory has already been introduced in the discussion of the Ewens sampling theory for the infinite alleles model. This theory shows that the sample size and the number of distinct alleles observed in the sample are sufficient to give an expected configuration of allele counts. From the observed and expected configurations, a number of tests might be

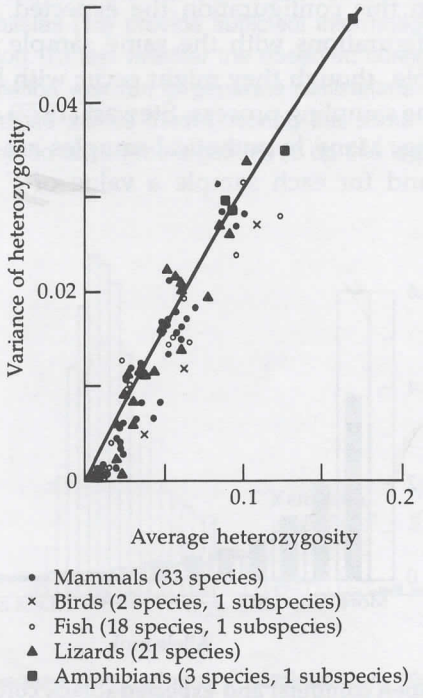


Figure 17. Mean and variance of heterozygosity among genes in vertebrates. The solid line is the theoretical curve derived by Nei, Fuerst, and Chakraborty for the infinite alleles model when the mutation rate to neutral alleles varies among genes in such a manner that the variance in mutation rate equals the square of the mean mutation rate. (After Nei et al. 1976.)

used to determine whether the observed sample fits the expected values of the model. The idea of directly testing the fits of the distributions by a χ^2 test has intuitive appeal, but apart from an informal heuristic consideration (Clark 1987b), this test implicitly assumes a multinomial sampling from a fixed distribution and does not permit the more complex sampling that goes into the infinite alleles model. Figure 18 shows histograms of the observed and expected allele frequency configurations for alleles in a human population defined by a restriction fragment length polymorphism. In this particular example, there appears to be a slight excess of the common allele, which is consistent with a number of causes of departure from the infinite alleles model (see Box F).

Rather than testing the fit of the whole allele configuration, Watterson (1978) showed that the *F* statistic, defined as the sum of the squared allele frequencies (adjusted for sampling) serves as a good test statistic, provided the confidence interval on the expected *F* can be generated. This requires a means of generating many realizations of the process of infinite alleles sampling. It is true that there is an expected allele frequency configuration under the model, and from this configuration the expected *F* can be calculated. However, other configurations with the same sample size and number of alleles are also possible, though they might occur with lower probability. By actually simulating the sampling process, Stewart (1977) devised an algorithm widely used in testing. Many hypothetical samples are generated using the Stewart algorithm, and for each sample a value of *F* is calculated. If the

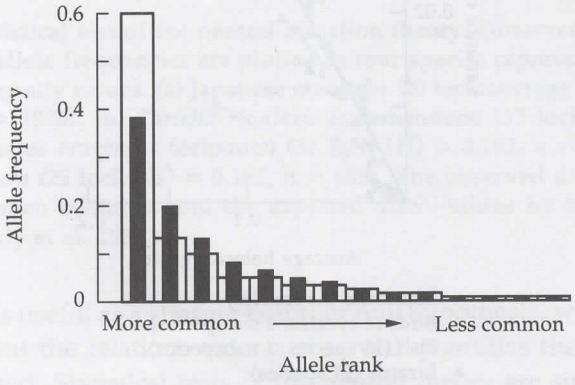


Figure 18. Observed (open columns) and expected (black columns) allele frequency distribution of the *HRAS-1* locus in humans, identified by Southern blotting with the *pLM*0.8 probe and *Taq*I digests. Observed data are from Baird et al. (1986), and expected distribution was generated using Ewen's sampling theory. In this sample of 490 genes there were 14 distinct alleles, 4 of which were present in just one individual. (From Clark 1987b.)

observed F falls within a tail of this distribution (say, the top or bottom 2.5%), then we would conclude that the probability that the sample came from a population that obeys the infinite alleles model is less than 5%. An example of this test is given in Box F.

The Ewens-Watterson Test

F

Keith et al. (1985) obtained a sample of *Drosophila pseudoobscura* from the Gundlach-Bundschu Winery in Sonoma Valley, California, and isolated a set of 89 chromosomes by a procedure similar to that depicted in Figure 5. Homogenized tissue from these 89 lines was subjected to sequential electrophoresis (a sensitive means of detecting charge and conformation differences among the protein products) and stained to reveal differences in xanthine dehydrogenase (*Xdh*) mobility. The researchers obtained a common allele that was present in 52 of the lines, one allele that was present in 9 lines, one allele that was present in 8 lines, two alleles present in 4 lines each, two alleles present in 2 lines each, and 8 singleton (or unique) alleles.

To see if this configuration of alleles fits the Ewens sampling distribution predicted by the infinite alleles model, first note that Ewens showed that the sample size (89) and the number of distinct alleles (15) provide sufficient information to determine the expected allele configuration. To test whether the observed configuration fits this expectation, a computer simulation was run to generate realizations of samples from populations that obey the infinite alleles model, having the same number of alleles and sample size as the observed data. [The algorithm to do this is described by F. Stewart

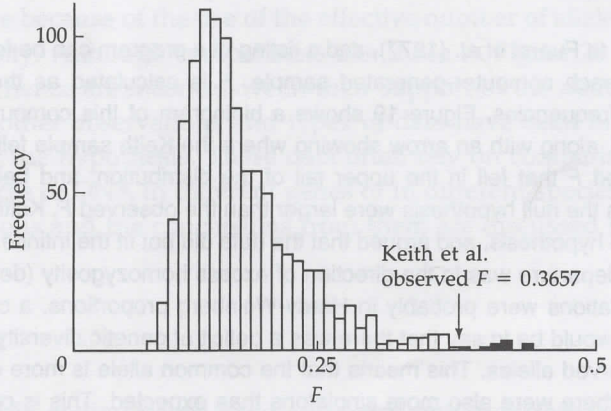


Figure 19. Computer-generated distribution of *F* obtained from 1000 samples from a population obeying assumptions of the infinite alleles model with $k = 15$ alleles and a sample of size $n = 89$ (as in the *Xdh* data from a sample of *Drosophila pseudoobscura* from the Gundlach-Bundschu Winery studied by Keith et al. 1985). The mean of *F* from the simulation was 0.1677, which is well below the observed *F* of 0.3657. A significant departure of the observed *F* from the predictions of the model is noted by the small tail area under the distribution to the right of the arrow.

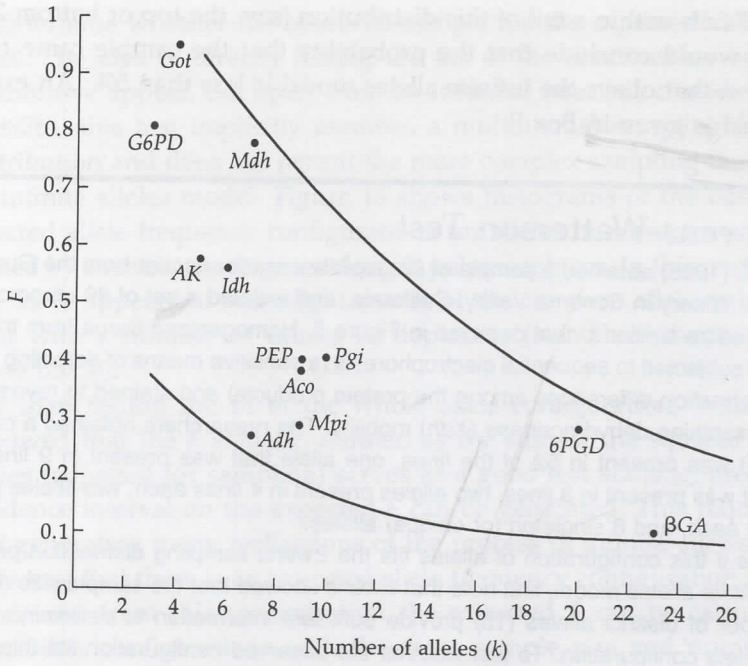


Figure 20. Gene identity (*F*) plotted against the observed number of alleles in a sample of 279 *E. coli*. The solid lines represent the upper and lower boundaries of the 95% confidence region, and the observation that all of the tested loci fall within these limits suggests good concordance with the infinite alleles model of neutral mutation. (From Whittam et al. 1983.)

in the Appendix to Fuerst et al. (1977), and a listing of a program can be found in Manly (1985).] From each computer-generated sample, F is calculated as the sum of the squared allele frequencies. Figure 19 shows a histogram of this computer-generated distribution of F, along with an arrow showing where the Keith sample fell. The sample had an observed F that fell in the upper tail of the distribution, and because so few values of F from the null hypothesis were larger than the observed F, Keith et al. (1985) rejected the null hypothesis, and argued that the data did not fit the infinite alleles model very well. The departure was in the direction of excess homozygosity (deficit of H), but since the populations were probably in Hardy-Weinberg proportions, a clearer way to state the result would be to say that there was a deficit of genetic diversity for the given number of observed alleles. This means that the common allele is more common than expected, and there were also more singletons than expected. This is consistent with purifying selection: selection acting to eliminate the rare, slightly deleterious alleles that continually enter the population by mutation. It is also consistent with an historical effect in which many alleles were previously lost and the population has not yet returned to equilibrium.

The results of the Ewens-Watterson test can also be reported graphically, as in Figure 20. Each gene yields a point specified by the number of distinct alleles and the

observed *F*. The two curves represent the 95% confidence interval generated by the Ewens-Watterson sampling theory. A quick check of the concordance of the data with the model can be made by seeing whether points remain in this confidence region. Although *Xdh* in *Drosophila pseudoobscura* provides a dramatic departure from the infinite alleles model, results like those plotted in Figure 20, which show an acceptable fit to neutrality, are more commonly obtained. (Data from Whittam et al. 1983).

One of the predictions of the neutral theory is that the ratio of the variance in the number of substitutions in a lineage to the mean number of substitutions should be close to one, because mutations accumulate approximately by a Poisson process (Kimura 1983). However, when Gillespie (1986a) calculated this ratio using DNA sequences from four nuclear genes and five mitochondrial genes, he found that the ratio varied from 0.16 to 35.55. This range includes values that differ with high statistical significance from the neutral expectation of 1. Although the neutral theory does not predict that nucleotide substitution is a perfect Poisson process, the range in the variance:mean ratio is much greater than would be expected. Gillespie explains the difference by saying that the process of substitution appears to be episodic, and in other publications (Gillespie 1978, 1986b) he presented models that reproduce the Ewens sampling distribution of allele configurations in the face of natural selection. The point is that, although the neutral theory is able to explain a great deal about molecular evolution, there are some issues that remain subjects of intense debate.

The above discussion is a preliminary introduction to Chapter 7, which is relevant here because of the use of the effective number of alleles and average heterozygosity. Although comparisons like those in Figure 20 are lacking in power, and hence are inconclusive in their support of the neutrality hypothesis, many other observations and types of data have been brought to bear in assessing the hypothesis. These data often rely on comparison of nucleotide sequences of DNA in different genes or in different species. These types of comparisons and the conclusions from them are discussed in Chapter 7.

PROBLEMS

1 Most protein-coding genes have a forward mutation rate (normal to mutant) that is at least an order of magnitude greater than the reverse mutation rate (mutate back to normal). Why should this be the case?

2 A classical bacterial experiment demonstrated that mutations occur at random and not in response to specific selection pressures for them. The experiment used sterilized velvet to imprint the geometrical pattern of

bacterial colonies on an agar surface in a petri dish (a "plate"), which was used to replicate the pattern by impressing the velvet on sterile nutrient agar in a selective plate containing an antibiotic. Colonies on the original plate giving resistant cells on the selective plate were dispersed into single cells, spread onto a nutrient agar plate without antibiotic, and allowed to multiply into colonies. This procedure was repeated until one or more colonies on the unselective media consisted exclusively of antibiotic-resistant cells. How does this experiment prove the point?

3 How many transition and transversion mutations are possible? Given that the observed ratio is about 2:1, how frequently do transitions occur relative to their expected value?

4 Estimation of mutation rates from bacterial cultures can be tricky because, if a mutation occurs early in the life of a culture, the final frequency will be very high; but if it occurs late, the final frequency will be low. The fluctuation test is a method for getting around this problem by growing many smaller cultures and estimating the mutation rate from the proportion of cultures that contain no mutations using the zero term of the Poisson distribution $P_0 = \exp(-\mu N)$ where P_0 is the proportion of cultures with no mutations, μ is the mutation rate, and N is the average number of cells per culture. In one experiment for bacteriophage T1 resistance, 11/20 cultures contained no mutations and the average number of cells per culture was 5.6×10^8. Estimate μ.

5 If recessive lethals occur independently in *Drosophila* autosomes, and the probability that an autosome contains one or more recessive lethals is 0.35 (a typical figure for chromosomes isolated from natural populations), what is the average number of recessive lethals per chromosome? Assume that the distribution of lethals is Poisson so that the probability of a chromosome containing exactly i lethals is $P_i = (m^i/i!)\exp(-m)$, where m is the mean.

6 The doubling dose of radiation is the quantity of radiation that induces as many mutations as occur spontaneously, so the total mutation rate of organisms exposed to the doubling dose equals two times the spontaneous mutation rate. Below are the induction rates per rad of x-rays (a standard measure of dose) for various genetic end points in irradiated male mice, along with the spontaneous rates. What are the corresponding doubling doses?

	Induction rate/rad	**Spontaneous rate**
Dominant lethals	5×10^{-4}/gamete	$2 - 10 \times 10^{-2}$/gamete
Recessive visibles	7×10^{-8}/locus	8×10^{-6}/locus
Reciprocal translocations	$1 - 2 \times 10^{-5}$/cell	$2 - 5 \times 10^{-4}$/cell

7 For irreversible mutation with a forward mutation rate $\mu = 5 \times 10^{-6}$,

calculate the allele frequency p after 10, 100, 1000, and 10000 generations, assuming $p_0 = 1.0$.

8 If a transposable genetic element becomes fixed at a particular site but undergoes deletion at the rate of one percent per generation, how many generations are required to decrease the frequency of the element at the site to 90%?

9 The following data give the frequency q of bacteria resistant to a bacteriophage after t generations of chemostat growth. At $t = 12$ hours a novel metabolite was added to the medium.

a. What is the basal rate of mutation to bacteriophage resistance?

b. What is the effect of the novel metabolite on the mutation rate?

t	q	t	q
0	1×10^{-6}	16	7.04×10^{-6}
4	3×10^{-6}	20	7.08×10^{-6}
8	5×10^{-6}	24	7.12×10^{-6}
12	7×10^{-6}		

10 In the forward and reverse mutation model, what is the equilibrium frequency p of A if

a. $\mu = 10^{-5}$ and $\nu = 10^{-6}$?

b. μ is increased tenfold?

c. ν is increased tenfold?

d. both μ and ν are increased tenfold?

11 In the forward and reverse mutation model, show that the time required for the allele frequency to go half way to equilibrium is approximately $t = 0.7/(\mu + \nu)$ generations. (Hint: Use the approximation that $\ln(1 - x) \approx -x$ when x is small.) What time is required to go halfway to equilibrium when $\mu = 10^{-5}$ and $\nu = 10^{-6}$?

12 In the irreversible mutation model, what is the frequency q_t of allele a in generation t if the mutation rate changes from generation to generation? If the equation $q_t = q_0 + \mu t$ is applied to this situation, what value corresponds to μ?

13 Calculate the autozygosity F after 200 generations in a random mating population of effective size $N = 50$.

14 In an isolated random mating population of effective size N, how many generations of random genetic drift are required to produce the same average inbreeding coefficient F as obtained in one generation of brother-sister mating (for which $F = 1/4$)? Use the approximation $[1 - 1/(2N)]^t \approx \exp(-t/2N)$.

15 Suppose a gene has eight alleles at frequencies 0.55, 0.20, 0.09, 0.06, 0.04,

0.03, 0.02, and 0.01. What is the effective number of alleles? What would the effective number be if each allele had a frequency of 0.125?

16 Why is the effective number of alleles essentially independent of the number of rare alleles?

17 What is the expected number of neutral alleles in a sample of size $n = 10$ alleles when $\theta = 4N\mu = 0.1$, 1, and 10, respectively?

18 What is the equilibrium heterozygosity in a population of effective size 50 if new neutral mutations are introduced at a rate 10^{-5} by mutation and at a rate 10^{-3} by migration?

19 Suppose $\theta = 4N\mu = x/(1 - x)$ where $0 < x < 1$. Show that the expected number of neutral alleles in a sample of size $n = 2$ alleles is $1 + x$.

20 If the average number of alleles of a gene is $1 + x$ per diploid individual, where $0 < x < 1$, then what is the heterozygosity? (Note that one diploid individual is a random sample of two alleles, and see Problem 19.)

21 Suppose a sample of $\sqrt{N}$ alleles is chosen at random from a population of effective size N. What is the expected value of the most recent time that the ancestry of the alleles coalesces into fewer than $\sqrt{N}$ lineages? What is the answer for a sample of 8 genes in a population of effective size 60?

Chapter Four

Natural Selection

The theory of evolution by means of natural selection is based on a number of premises and the conclusions that follow from them. The first premise is that all species have more offspring than can possibly survive and reproduce; this premise follows from the observation that, under optimal environmental conditions, population growth tends to be exponential. The second premise is that organisms vary in their ability to survive and reproduce, a postulate whose validity is obvious from the simple observation that some organisms do survive and reproduce and others do not. The third premise is that part of the variation in ability to survive and reproduce is hereditary; this premise is a key one, and Darwin took great pains in the *Origin* to document the occurrence of heritable variation affecting virtually all traits in natural populations and in populations of domesticated species.

The premise of genetic variation can be stated in modern terminology by saying that certain genotypes are more adapted to their environment than other genotypes, as evidenced by their greater **fitness** (their greater ability to survive and reproduce in that environment). The conclusion from the three premises is the occurrence of **natural selection**: the process by which genotypes with greater fitness leave, on the average, more offspring than do less fit genotypes. Because of natural selection, favorable alleles that promote higher fitness will be overrepresented in the next and succeeding generations. As these favorable alleles increase in frequency in the population, more of the fitter genotypes are formed. As a result, the types and frequencies of alleles in the population gradually change so as to promote greater adaptation to the environment.

DEFINITIONS OF FITNESS

Perhaps the best way to model the process of natural selection is to consider all of the stages in the life cycle of an organism. A diagram of a life cycle is presented in Figure 1. Beginning with the zygotic stage, an organism must undergo growth and development to become a mature adult. During this time, the different genotypes may have different likelihoods of success, and selection has an opportunity to act by differential **viabilities** of the genotypes. There are many complications that can be added to the viability component of selection, including effects of relative frequencies of genotypes, group interactions, and environmental effects. Selection can occur among a set of mature adults by differential mating success. This component of selection is known as **sexual selection**. Darwin, in his book *The Descent of Man and Selection in Relation to Sex*, envisaged sexual selection as occurring by either of two mechanisms. By the first mechanism, mating success may be determined in part by choice of potential mates (female choice, male choice, or some kind of interaction). The other mechanism of sexual selection is direct competition between members of the same sex for access to the opposite sex. An example of this is the butting contests held by bighorn sheep. These contests are not actually fights for survival, since the loser rarely incurs fatal injuries, but they are fights for the survival of the victor's genes. Sexual selection can be a property of the individual diploid genotype, since in some cases one sex determines the outcomes of matings; in other cases, the formation of mating pairs is determined by both sexes. In the latter situation, a model of sexual selection would have to have parameters specifying the mating probability for each possible mating pair, hence this component of selection must be a property of pairs of genotypes. After matings have occurred, there is opportunity for selection if the array of offspring genotypes differs from that predicted by simple Mendelian segregation. Deviations could be due to differential production of gametes by heterozygotes, or differential success of gametes in accomplishing fertilization. The former is

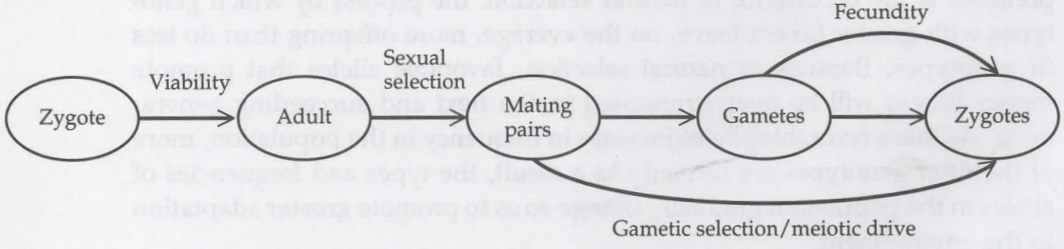


Figure 1. A generalized life cycle showing the components of selection.

known as **meiotic drive**, the latter as **gametic selection**. In both cases, selection affects the haploid gamete. Selection can also occur by different mating types producing different numbers of offspring, a component known as **fecundity selection**.

The above discussion relates to organisms such as humans and *Drosophila* that have a conspicuous, long-lived diploid stage, and a much shorter stage as haploid gametes. Other organisms may have a longer haploid stage, or even a predominant haploid stage, with a short and relatively inconspicuous diploid stage. Many bizarre life histories can be found in the plant kingdom and among marine invertebrates. In these cases it is necessary to tailor the models to the specific case.

COMPLICATIONS IN MEASURING FITNESS

Given the premises of evolution described by Darwin — reproductive excess, reproductive differences among individuals, and heritability — it is clear that a primary experimental test of the theory is to find cases where different, identified genotypes have different fitnesses. The problem of fitness estimation is difficult for several reasons. One of the biggest obstacles is the biological complexity of fitness. An allelic substitution may have pleiotropic effects at any of the stages discussed above, so the first experimental problem is to isolate the components of fitness. It is critical that the components be isolated and tested separately, because when pooled effects are measured over the period of a complete generation, the apparent fitness estimates could spuriously confound effects of individual components. The reason, as we will see below, is that different fitness components have different effects on the dynamics and equilibrium properties of allele frequencies. A number of different strategies have been used for estimating the components of fitness, and these can be split broadly into two classes: laboratory manipulations, and field sampling of different age classes (Manly 1985). In both cases it is critical to design the sampling scheme to suit the biology of the organism.

In addition to the biological difficulty in identifying relevant life stages for measuring fitness components, another problem is the genetic complexity of fitness. It is obvious that the probability of surviving or the chance of successfully attracting mates can involve genetically complex traits, including many genes. The question is whether the marginal effects of one or a few identified genes are important. In the case of lethal or nearly lethal genes, the effects of a single locus on fitness are obvious, but if lethal alleles are rare, then they may account for relatively little of the total variation in fitness.

A final major obstacle to obtaining useful fitness estimates is statistical. Selection coefficients as small as a 1% difference in viability or fecundity can

have a major impact on the fate of a gene, so experiments should be designed to be able to detect selection this weak. Experiments that fail to detect fitness differences among genotypes are not illuminating if they have inadequate sample sizes and cannot detect selection differences smaller than 10%. On the other hand, failure to detect selective differences in experiments that could detect selection coefficients as small as 0.5% (Dykhuizen and Hartl 1983b) allow the conclusion that the genotypes really are quite similar in fitness, though not necessarily selectively neutral.

In the 1970s many investigators were concerned with finding evidence either in favor of or in contradiction to the neutral theory of molecular evolution. In particular, the issue was whether the allozyme variation almost universally seen with electrophoresis was actively maintained by natural selection, or whether it was retained by a combined process of mutation and drift. This debate gave impetus to attempts to estimate the fitness of allozyme genotypes. As with hemlines, however, there are trends in science. Currently there is much less effort to obtain estimates of the effects on fitness of allelic substitution. This does not mean that the problem is solved, or that it is no longer interesting. Investigators have realized the magnitude of the difficulties, and the advent of molecular techniques have opened many new avenues of research.

In particular, problems relating to evolution at the molecular level have become much more popular areas of investigation. Molecular approaches result in statements about natural selection without actually measuring fitness components: both significant departures from predictions made by models of neutral mutations and marked conservation of sequences have been interpreted as evidence of selection. But these conclusions lack the appeal of direct estimates of fitness in not specifying a mechanism of selection, and in not providing estimates of parameters that allow population prediction. In this sense, the molecular approaches are more retrospective, while the approach of estimating selection components is prospective (Ewens 1979).

Table 1. Assumptions of constant selection model.

(1) Diploid organism
(2) Sexual reproduction
(3) Nonoverlapping generations
(4) Infinite population size
(5) Random mating
(6) No migration
(7) No mutation
(8) Selection occurs, fitnesses constant and equal in the two sexes

CLASSICAL MODELS (VIABILITY)

One Gene with Two Alleles

Despite the many complexities of fitness, the consequences of selection are most conveniently explored using the simple model with assumptions outlined in Table 1. The formulation of this model appears in Table 2. The model is essentially the Hardy-Weinberg model, but now differential survival of genotypes is permitted. Selection is assumed to occur in the diploid genotypes, not in the gametes, and segregation is assumed to be Mendelian. How allele frequencies change during one generation of selection is shown in Table 2. The gametic frequencies of A and a in generation t are p_t and q_t (denoted p and q for brevity). Genotype frequencies in zygotes are given by the Hardy-Weinberg principle (see Chapter 1), and the relative fitnesses (viabilities) of AA, Aa, and aa are denoted w_{11}, w_{12}, and w_{22}, respectively. By definition, zygotes survive in the ratio $w_{11} : w_{12} : w_{22}$, so the ratio of $AA : Aa : aa$ among adults is

$$p^2w_{11} : 2pqw_{12} : q^2w_{22}$$

The ratio of $A : a$ in gametes is therefore

$$[p^2w_{11} + 1/2(2pqw_{12})] : [1/2(2pqw_{12}) + q^2w_{22}]$$

(The 1/2's enter the ratio because Aa heterozygotes produce 1/2 A and 1/2 a gametes due to Mendelian segregation.) The ratio of $A : a$ in gametes readily simplifies to

$$p(pw_{11} + qw_{12}) : q(pw_{12} + qw_{22})$$

To obtain the gametic frequencies, these terms must be **normalized** so that the frequencies sum to 1. Dividing each value in the above ratio by the overall sum accomplishes this. The sum is

$$p(pw_{11} + qw_{12}) + q(pw_{12} + qw_{22}) = p^2w_{11} + 2pqw_{12} + q^2w_{22} \qquad \textbf{(4.1)}$$

This sum, usually denoted $\overline{w}$, is simply the **average fitness** of the population. Dividing the gametic ratios by $\overline{w}$ gives the gametic frequencies in generation $t + 1$, namely

$$p_{t+1} = p(pw_{11} + qw_{12})/\overline{w} \qquad \textbf{(4.2a)}$$

and

$$q_{t+1} = q(pw_{12} + qw_{22})/\overline{w} \qquad \textbf{(4.2b)}$$

which are denoted p' and q' in Table 2. Thus, in one generation of selection, the allele frequency of A changes from p to

Table 2. Change in allele frequency over one generation of selection.[a]

Generation	Life-history stage	Frequencies		
t	Gametes	A		a
	Gametic frequencies ($p + q = 1$)	$p_t(= p)$		$q_t(= q)$
$t + 1$	Genotypes	AA	Aa	aa
	Genotype frequencies (zygotes)	p^2	$2pq$	q^2
	Fitness (viability)	w_{11}	w_{12}	w_{22}
	Ratio of genotypes (adults)	p^2w_{11} :	$2pqw_{12}$:	q^2w_{22}
$t + 1$	Gametes	A		a
	Ratio of gametic frequencies	$p^2w_{11} + (1/2)(2pqw_{12})$	:	$(1/2)(2pqw_{12}) + q^2w_{22}$
		$p(pw_{11} + qw_{12})$	:	$q(pw_{12} + qw_{22})$
	Average fitness ($\bar{w}$)	$\bar{w} = p(pw_{11} + qw_{12}) + q(pw_{12} + qw_{22})$		
		$= p^2w_{11} + 2pqw_{12} + q^2w_{22}$		
	Gametic *frequencies* ($p' + q' = 1$)	A		a
		$p_{t+1} = p'$		$q_{t+1} = q'$
		$p' = p(pw_{11} + qw_{12})/\bar{w}$		$q' = q(pw_{12} + qw_{22})/\bar{w}$
	Change in allele frequency of A (Δp)	$\Delta p = p' - p = pq[p(w_{11} - w_{12}) + q(w_{12} - w_{22})]/\bar{w}$		

[a] The relative viabilities of *AA*, *Aa*, and *aa* are denoted w_{11}, w_{12}, and w_{22}, respectively.

$$p' = \frac{p(pw_{11} + qw_{12})}{\overline{w}} \tag{4.3}$$

It is often useful to know $p' - p$, often symbolized Δp, the difference in allele frequency caused by one generation of selection; in this case,

$$\Delta p = \frac{pq[p(w_{11} - w_{12}) + q(w_{12} - w_{22})]}{\overline{w}} \tag{4.4}$$

Numerical Example of Allele Frequency Change with Selection

A

At this point, a numerical example illustrating use of the above equations might be useful. Let us consider an extreme example in which genotypes *AA*, *Aa*, and *aa* survive in the ratio $w_{11} = 1$, $w_{12} = 0.5$, and $w_{22} = 0$ (i.e., *aa* is lethal, and *Aa* has a probability of survival only half as great as *AA*). Suppose that initially the allele frequency of *A* among gametes is $p = 0.1$. What will the value of p be among gametes in the next generation? (It is natural to wonder how the allele frequency of *A* could be a mere 0.1 considering the large amount of selection against *a*. Such a situation can easily be created experimentally, as the investigator can set the initial allele frequency at any predetermined value by appropriate choice of genotypes in the founding population; the situation can also arise in nature when a change of environment causes previously favored genotypes to become disfavored.) To calculate p among gametes in the next generation (p'), first calculate $\overline{w}$ as

$$\overline{w} = p^2w_{11} + 2pqw_{12} + q^2w_{22} = (0.1)^2(1) + 2(0.1)(0.9)(0.5) + (0.9)^2(0) = 0.10$$

then use

$$p' = p(pw_{11} + qw_{12})/\overline{w}$$

to obtain $p' = (0.1)[(0.1)(1) + (0.9)(0.5)]/(0.10) = 0.55$. In this case, $\Delta p = 0.55 - 0.1 = 0.45$, which could have been calculated directly from the equation

$$\Delta p = pq[p(w_{11} - w_{12}) + q(w_{12} - w_{22})]/\overline{w}$$

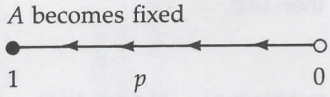
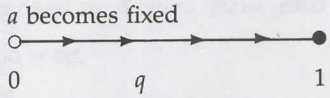
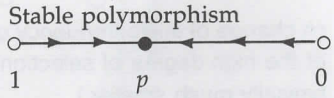
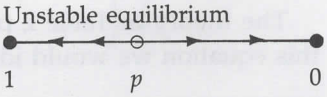
$$= (0.1)(0.9)[(0.1)(0.5) + (0.9)(0.5)]/(0.10) = 0.45$$

(A change of allele frequency of 0.45 in one generation of selection is extreme because of the high degree of selection in this example; values of Δp for realistic examples are typically much smaller.)

The theory in Table 2 provides an equation for p_{t+1} in terms of p_t. From this equation we would ideally wish to find an expression for p_t in terms of

p_0. Unfortunately, a general expression for p_t cannot be obtained for selection except in a few special cases. The long-term effects of selection when the relative fitnesses are constant can be determined rather easily, however; they are summarized in Table 3. The first two cases in Table 3 follow immediately from the expression for Δp in Table 2. First, note that there are always at least two equilibria (values of p_t such that $p_{t+1} = p_t$ or $\Delta p = 0$), namely $\hat{p} = 0$ and $\hat{p} = 1$. Note that the symbol $\hat{p}$ (pronounced "p hat") is used to indicate an equilibrium value of p. In the first instance, $\Delta p = 0$ because $p = 0$; in the second instance $\Delta p = 0$ because $q = 0$. Then, too, $\overline{w}$ must always be positive, and p and q must certainly not be negative. Therefore, if $w_{11} > w_{12} > w_{22}$, then Δp will always be positive because $p(w_{11} - w_{12}) + q(w_{12} - w_{22})$ will always be positive; but $\Delta p > 0$ means that p always increases, so p must eventually go to $p = 1$. Conversely, when $w_{11} < w_{12} < w_{22}$, as in case 2, then Δp must always be negative, so $p \rightarrow 0$ (implying that $q \rightarrow 1$). Examples of these two cases are shown in Figure 2: in panel a, $p \rightarrow 1$ for any initial frequency p_0 because $w_{11} > w_{12} > w_{22}$; in panel b, $p \rightarrow 0$ for any p_0 because $w_{11} < w_{12} < w_{22}$. Note in both cases that the most rapid changes in allele frequency occur when the allele frequency is around 0.5.

Table 3. Summary of behavior of the one-locus two-allele viability model.

	AA *Aa* *aa*	
Directional selection		
Case 1	$w_{11} > w_{12} > w_{22}$	*A* becomes fixed 
Case 2	$w_{11} < w_{12} < w_{22}$	*a* becomes fixed 
Overdominance		
Case 3	$w_{11} < w_{12} > w_{22}$	Stable polymorphism 
Underdominance		
Case 4	$w_{11} > w_{12} < w_{22}$	Unstable equilibrium 

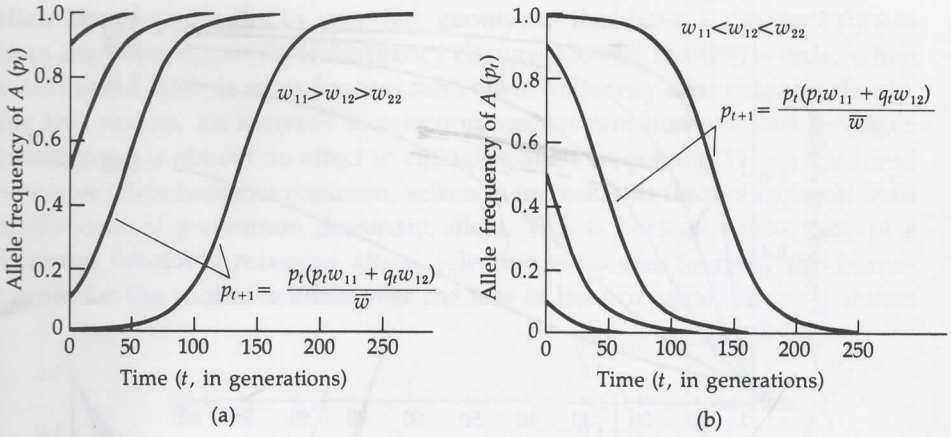


Figure 2. Changes of allele frequency against time when (a) allele *A* is favored, and (b) allele *a* is favored. (a) Fitnesses assumed are $w_{11} = 1$, $w_{12} = 0.9500$, $w_{22} = 0.9025$. (b) Fitnesses assumed are $w_{11} = 0.9025$, $w_{12} = 0.9500$, and $w_{22} = 1$. In both cases $w_{12} = \sqrt{w_{11}w_{22}}$, which allows an explicit expression for p_t to be obtained.

There is a special terminology to describe certain types of equilibria. Observe in Figure 2a that, when p is close to the equilibrium at $p = 0$, p moves farther away from the equilibrium in subsequent generations; such an equilibrium is said to be **unstable**. On the other hand, when p is close to the equilibrium at $p = 1$ in Figure 2a p moves closer to the equilibrium in subsequent generations; such an equilibrium is said to be **locally stable**. Actually, the equilibrium at $p = 1$ in Figure 2a is **globally stable**, which means that $p \rightarrow 1$ irrespective of the starting value of p. Any globally stable equilibrium must also be locally stable, but whenever there is more than one locally stable equilibrium, none will be globally stable. In Figure 2b, $\hat{p} = 0$ is globally stable, whereas $\hat{p} = 1$ is unstable.

Figure 3 shows the results of selection favoring the *A* allele when *A* is dominant to *a* (i.e., $w_{11} = w_{12}$ and $w_{12} > w_{22}$). The fitnesses of *AA*, *Aa*, and *aa* in this case are often written $1 + s{:}1 + s{:}1$ (with $s > 0$), where s is called the **selection coefficient.** This is the case when a mutation occurs that produces a rare favorable dominant allele. Note that a favorable dominant allele will immediately increase in frequency, even when it is rare. When the dominant allele becomes high in frequency, however, it changes in frequency much more slowly. Figure 4 shows the situation when the new mutation results in an allele *a* that is recessive to *A* (i.e., $w_{11} < w_{12}$ and $w_{12} = w_{22}$). When a recessive allele is rare, its initial rate of change in allele frequency is slow whether it is favored or disfavored. This is because most of the alleles in the population are in heterozygotes when an allele is rare. A rare recessive

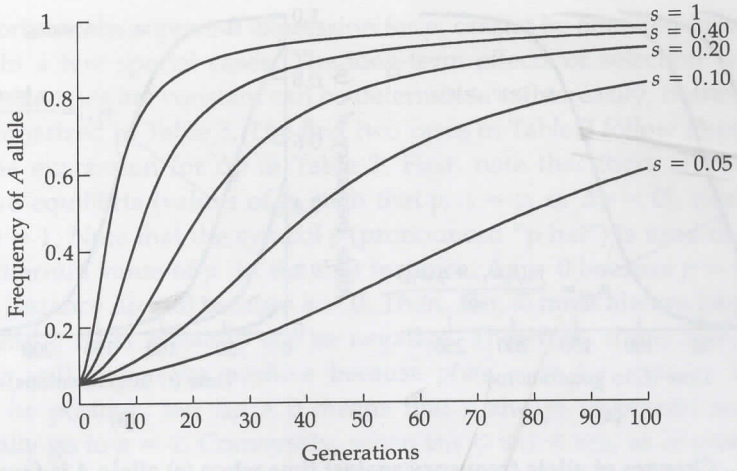


Figure 3. Directional viability selection favoring a new dominant mutation. The fitnesses are $w_{11} = w_{12} = 1 + s$ and $w_{22} = 1.0$. The selection coefficient (s, given for each curve) varies from 0.05 to 1.00. The curves show the frequency of allele A when introduced at an initial frequency of 0.05. A rare favorable dominant mutation increases in frequency rapidly, even if selection is fairly weak.

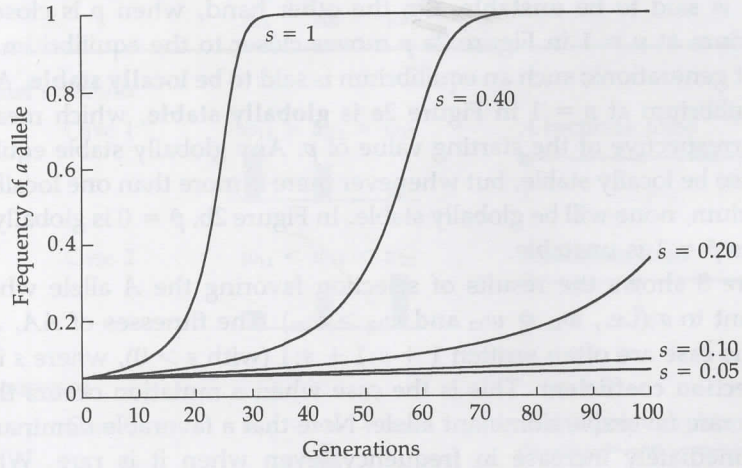


Figure 4. Directional viability selection favoring a new recessive mutation. The fitnesses are $w_{11} = w_{12} = 1$, and $w_{22} = 1 + s$. The selection coefficient (s, given for each curve) varies from 0.05 to 1.00. The curves show the frequency of allele a when introduced at an initial frequency of 0.05. A recessive mutation with only a slight advantage requires a long time to increase when it is initially rare. Once a recessive becomes common, progress to fixation is rapid.

allele therefore results in very few genotypes that have a different fitness from the majority, so allele frequency changes slowly. In other words, when a disfavored allele is recessive and rare, allele frequency changes very slowly. For this reason, an increase in selection against rare homozygous recessive genotypes has almost no effect in changing allele frequency. When a favored recessive allele becomes common, selection increases its frequency faster than in the case of a common dominant allele. This is because in the case of a common favorable recessive allele, selection continues to favor the homozygote for the recessive allele over the less fit heterozygote. Figure 5 shows

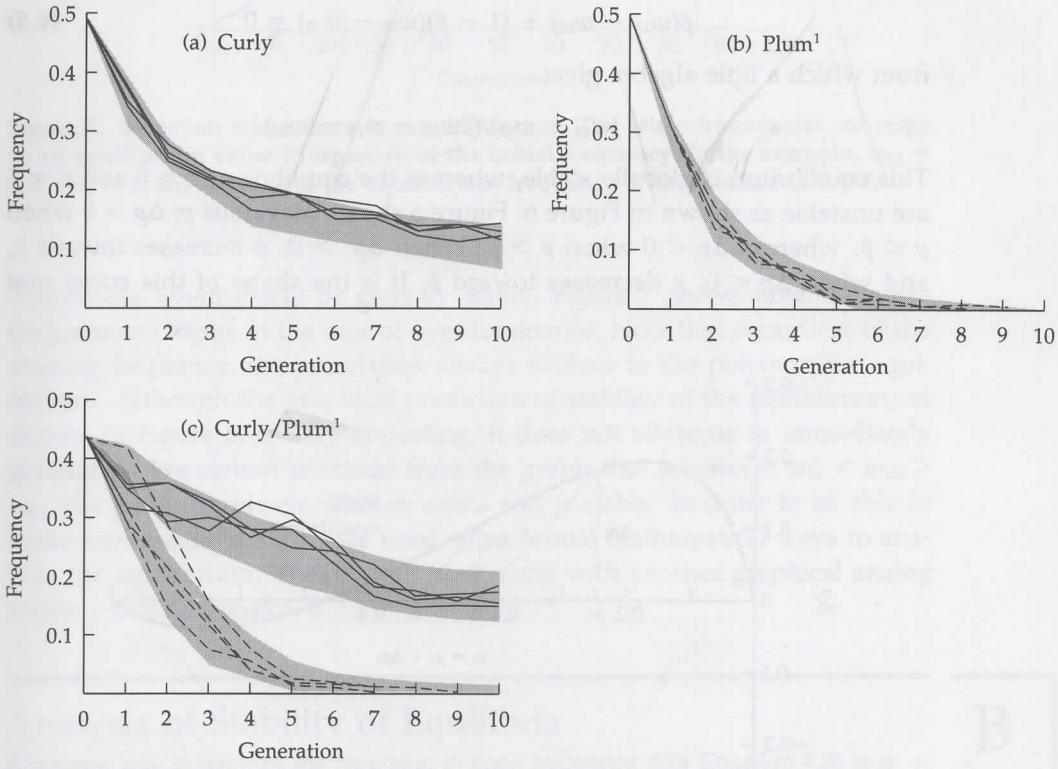


Figure 5. Trajectories of observed frequencies of balancer chromosomes in replicate populations of *Drosophila melanogaster*. (a) Only the *Curly* (*Cy*) marked second chromosome was segregating with a wildtype chromosome, so the only two genotypes are *Cy*/+ and +/+. (b) Only wildtype and *Plum*^{1} (Pm^1) chromosome present. From (a) and (b) it is evident that the *Cy*/+ genotype had a higher fitness relative to +/+ than the Pm^1/+ genotype. This was quantified by fitting an explicit model to these observed trajectories. (c) *Cy*, Pm^1, and wildtype segregating. Note that the presence of the Pm^1 resulted in a higher frequency of *Cy* through generation 6. Shaded areas are 95% confidence intervals of model fitted to the data. (From Clark et al. 1981.)

the observed course of selection involving recessive lethal balancer chromosomes in *Drosophila melanogaster*. The theoretical prediction fits rather well, considering the simple nature of the model.

Overdominance with one gene and two alleles

In case 3 of Table 3, $w_{11} < w_{12}$ and $w_{22} < w_{12}$, so the heterozygote is superior in fitness to both homozygotes, a condition known as **overdominance**. There is a third equilibrium in addition to $p = 0$ and $p = 1$ because $p(w_{11} - w_{12}) + q(w_{12} - w_{22})$ can equal 0. The third equilibrium allele frequency, $\hat{p}$, is found by solving

$$\hat{p}(w_{11} - w_{12}) + (1 - \hat{p})(w_{12} - w_{22}) = 0 \qquad \textbf{(4.5)}$$

from which a little algebra gives

$$\hat{p} = (w_{12} - w_{22})/(2w_{12} - w_{11} - w_{22})$$

This equilibrium is globally stable, whereas the equilibria at $p = 0$ and $p = 1$ are unstable as shown in Figure 6. Figure 6 shows Δp versus p; $\Delta p > 0$ when $p < \hat{p}$, whereas $\Delta p < 0$ when $p > \hat{p}$. When $\Delta p > 0$, p increases toward $\hat{p}$, and when $\Delta p < 0$, p decreases toward $\hat{p}$. It is the shape of this curve that

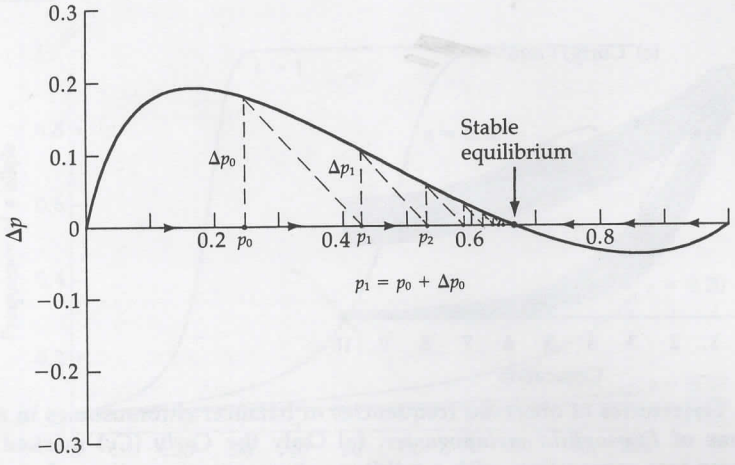


Figure 6. The change in allele frequency Δp plotted as a function of allele frequency p for the overdominant case ($w_{11} = 0.6$, $w_{12} = 1$, $w_{22} = 0.2$). Starting at allele frequency p_0, the height of the curve indicates that the allele frequency would change to $p_0 + \Delta p_0$ in the next generation. From this frequency, $p_1 = p_0 + \Delta p_0$, the change in allele frequency is Δp_1, giving $p_2 = p_1 + \Delta p_1$. This process is continued until the population arrives at the equilibrium point $\hat{p}$. The same procedure can be followed for allele frequencies greater than $\hat{p}$, showing again that the allele frequency converges to $\hat{p}$.

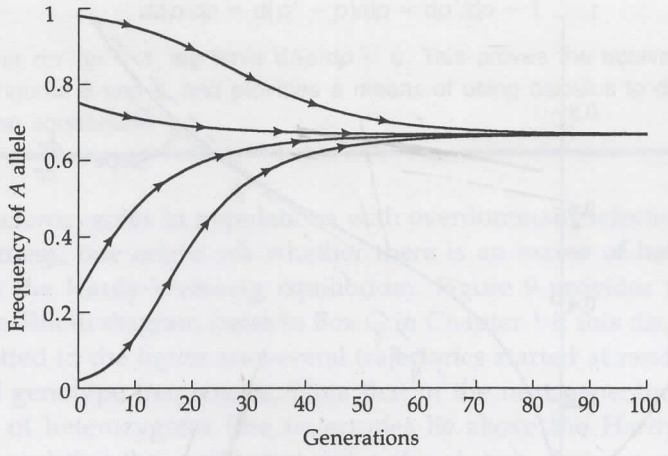


Figure 7. Selection when there is overdominance. The allele frequencies converge to an equilibrium value irrespective of the initial frequency. In the example, $w_{11} = 0.9$, $w_{12} = 1$, $w_{22} = 0.8$, and the equilibrium frequency of the A allele is 0.667.

determines when $\hat{p}$ will be globally stable. Figure 7 shows how the allele frequencies change in the case of overdominance. Note that regardless of the starting frequency, the population always evolves to the polymorphic equilibrium. Although the graphical prediction of stability of the equilibrium, as shown in Figure 6, is very appealing, it does not allow us to immediately generalize. We cannot conclude from the graph that whenever $w_{11} < w_{12} > w_{22}$, the polymorphic equilibrium exists and is stable. In order to be able to make statements like this, we need some formal mathematical ways to analyze the equilibrium. This is presented along with another graphical analog to the procedure in Box B.

Analysis of Stability of Equilibria

B

A general way to express the recursion in gene frequency (like Equation 4.3) is $p' = T(p)$, where T is an arbitrary function that transforms p to its value in the next generation. Figure 8 shows a hypothetical function $T(p)$. Suppose the point $\hat{p}$ is an equilibrium point, and we want to determine whether it is a stable point. One practical definition of a stable point is that when we start at a point $\hat{p} + \epsilon$, then the population would eventually return to $\hat{p}$. This means that $T(\hat{p} + \epsilon)$ is closer to $\hat{p}$ than was $\hat{p} + \epsilon$. We could of course test this numerically, but a more elegant (and accurate) way to determine stability makes use of the Taylor approximation of T. The Taylor approximation can be written as

$$T(\hat{p} + \epsilon) = T(\hat{p}) + \epsilon \frac{dT(\hat{p})}{dp} + \frac{\epsilon^2}{2!} \frac{d^2T(\hat{p})}{dp^2} + \ldots \approx T(\hat{p}) + \epsilon \frac{dT(\hat{p})}{dp}$$

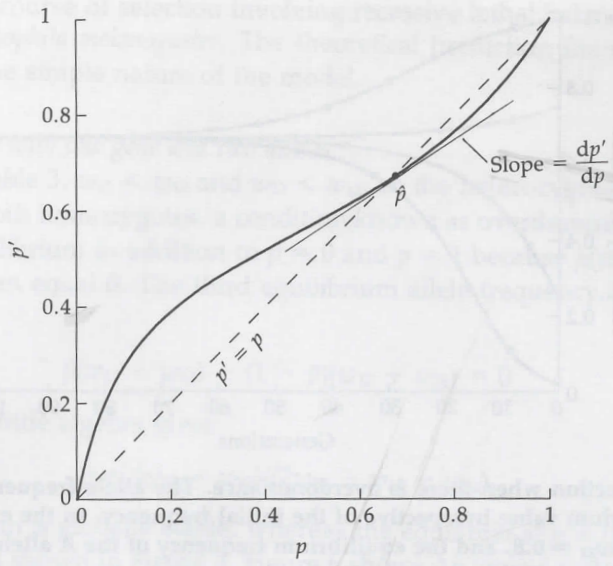


Figure 8. Another graphical means of determining stability of the equilibrium with overdominant viability selection. Here the fitnesses are $w_{11} = 0.6$, $w_{12} = 1$, and $w_{22} = 0.2$. The slope at the equilibrium point is $dp'/dp|_{p=\hat{p}}$, and this is the eigenvalue.

This approximation, known as a **linear approximation** because we simply consider the tangent along T as an estimate for p', is good for ϵ sufficiently small. Observe that $T(\hat{p}) = \hat{p}$, so the Taylor approximation tells us that p' is simply $p + \epsilon$ times the slope of T at the equilibrium. Thus $T(\hat{p} + \epsilon)$ is closer to $\hat{p}$ than was $\hat{p} + \epsilon$ if

$$\frac{dT(\hat{p})}{dp} < 1$$

and this is precisely the condition for stability of the equilibrium $\hat{p}$. Similarly, the equilibrium is unstable if the derivative is greater than 1. If the derivative is precisely 1, then the linear approximation is not adequate to determine the stability, and other methods must be used.

The stability condition can be seen graphically in Figure 8. The slope at the equilibrium point is dp'/dp, and if this slope is less than one, then for points above the equilibrium, $p' < p$, and for frequencies below the equilibrium, $p' > p$. Thus the condition $dp'/dp < 1$ indicates stability of the equilibrium, just as we showed algebraically. The value dp'/dp is called the **eigenvalue** of the recursion, and when it is evaluated at the various equilibria, it indicates the stability of the points. Sometimes eigenvalues are complex numbers or even pure imaginary numbers, indicating damped, divergent, or stable cycling dynamics, but these behaviors are encountered rarely in population genetic models.

To see that the condition $dp'/dp < 1$ is the same as the condition $d\Delta p/dp < 0$ (obtained from Figure 6), note that

$$d\Delta p/dp = d(p' - p)/dp = dp'/dp - 1$$

So whenever $dp'/dp < 1$, we have $d\Delta p/dp < 0$. This proves the equivalence of the graphs in Figures 6 and 8, and provides a means of using calculus to determine the stability of an equilibrium.

Since heterozygotes in populations with overdominant selection have the highest fitness, one might ask whether there is an excess of heterozygotes relative to the Hardy-Weinberg equilibrium. Figure 9 provides the answer using a De Finetti diagram (refer to Box G in Chapter 1 if this diagram is not clear). Plotted in the figure are several trajectories started at random combinations of genotype frequencies. Note that in the next generation, there is an excess of heterozygotes (the trajectories lie above the Hardy-Weinberg parabola), and that the equilibrium point also shows a heterozygote excess. Unfortunately, as we saw in Chapter 1, testing the goodness-of-fit to Hardy-Weinberg proportions is a poor test for selective differences, because it lacks statistical power, and several other factors could account for a departure.

Overdominance frequently leads to polymorphism in the sense discussed earlier in connection with electrophoresis (Chapter 1); that is to say, overdominance frequently leads to an equilibrium in which $\hat{p}$ is between 0.01 and 0.99. For this reason, overdominance has been much discussed as a possible

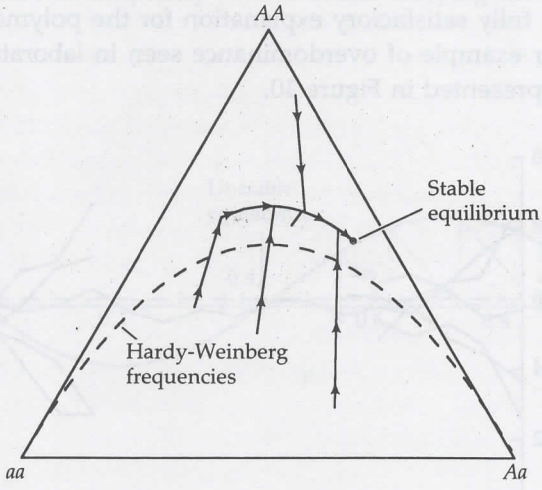


Figure 9. Trajectories of genotype frequencies plotted on a De Finetti diagram with $w_{11} = 0.8$, $w_{12} = 1$, and $w_{22} = 0.6$. The dashed line indicates the Hardy-Weinberg genotype frequencies, and several random starting points are seen to converge to the stable equilibrium, indicated by the dot. Note that there is a slight excess of heterozygotes at the equilibrium.

explanation for such polymorphisms as those involving allozymes. On the other hand, few proven cases of overdominance are known. The best known case involves two alleles coding for the β chain of human hemoglobin: $Hb\beta^A$, the normal allele, and $Hb\beta^S$, the allele leading to an amino acid substitution that causes sickle-cell anemia. Homozygous $Hb\beta^S/Hb\beta^S$ individuals are severely anemic due to sickling of red blood cells; $Hb\beta^A/Hb\beta^A$ individuals have normal hemoglobin but are sensitive to the type of malaria caused by the mosquito-borne protozoan parasite *Plasmodium falciparum*; $Hb\beta^S/Hb\beta^A$ heterozygotes are mildly anemic but tend to be resistant to falciparum malaria, apparently because red blood cells infested with the parasite undergo sickling and are removed from circulation. Since there are more than two alleles at the β-globin locus, further discussion of this situation will be included in a later section on multiple alleles.

Although overdominance is capable of maintaining genetic polymorphisms involving two alleles (except perhaps in very small populations; see Robertson 1962), a consideration of the possibility that overdominance provides a general explanation for polymorphism must await analysis of selection with multiple alleles. On the other hand, there is one situation in which simple overdominance is probably of great importance; this situation involves the polymorphic paracentric inversions found in many species of *Drosophila*. As discussed in Chapter 1, these inversions constitute coadapted "supergenes," and they frequently exhibit heterozygote superiority (Dobzhansky 1970). As we shall see, however, even in this case heterozygote superiority does not give a fully satisfactory explanation for the polymorphism seen in nature. Another example of overdominance seen in laboratory populations of *Drosophila* is presented in Figure 10.

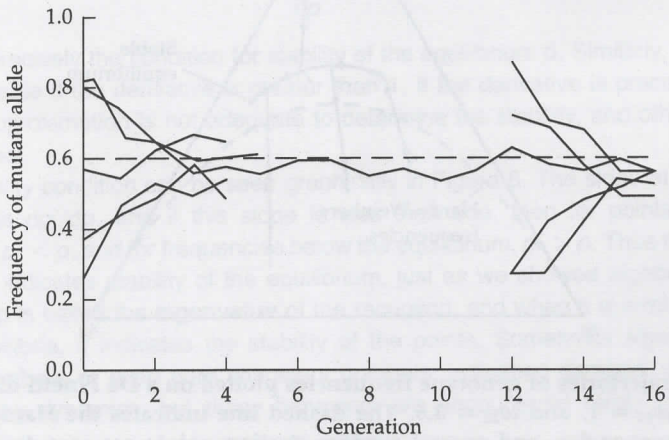


Figure 10. The frequency of a mutant allele when there is a heterozygous advantage after perturbations in generations 0 and 12. The broken line indicates the theoretical equilibrium. (From Prout 1971b.)

Heterozygote inferiority with one gene and two alleles

Case 4 in Table 3 is **heterozygote inferiority**, in which $w_{11} > w_{12}$ and $w_{22} > w_{12}$. Again there is an equilibrium at $\hat{p} = (w_{12} - w_{22})/(2w_{12} - w_{11} - w_{22})$, but in this case the equilibrium is unstable (Figure 11). The equilibria at $p = 0$ and $p = 1$ are both locally (but not globally) stable. When $p_0 < \hat{p}$, then $p \to 0$, whereas when $p_0 > \hat{p}$, then $p \to 1$. Trajectories of allele frequencies in the case of heterozygote inferiority are shown for several different starting conditions in Figure 12. Since the outcome of the selection depends on the initial allele frequency, this form of selection is sensitive to historical effects that may present local populations with different initial allele frequencies.

Heterozygote inferiority occurs primarily in conjunction with such chromosomal abnormalities as translocations, for which heterozygotes are semisterile due to the production of aneuploid gametes. Indeed, individuals carrying sufficiently complex chromosomal rearrangements produce no offspring in matings with normal individuals, because the offspring have gross chromosomal imbalances and are inviable; yet two individuals that each carry the chromosomal rearrangement can produce some viable offspring, though even in this case a substantial fraction of the offspring are inviable due to chromosomal imbalance. Such chromosomal rearrangements have been proposed as a possible method of population control, especially in insects (Foster et al. 1972), because any equilibrium involving both normal individuals and those carrying the rearrangement must necessarily be unstable; thus, for a sufficiently high initial frequency of individuals carrying the rearrangement,

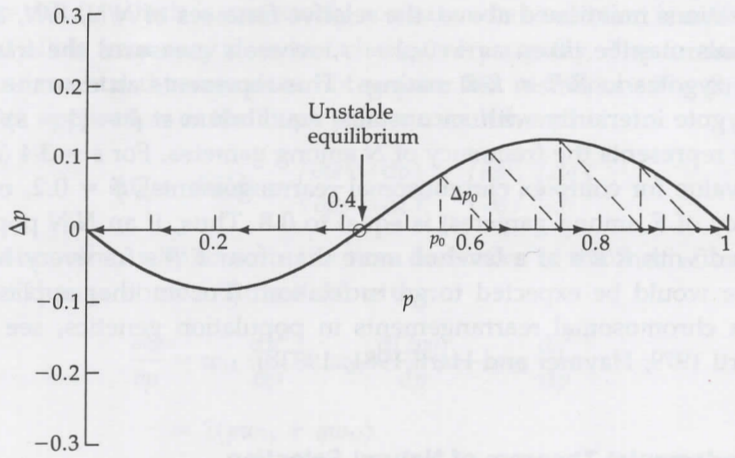


Figure 11. The change in allele frequency Δp plotted as a function of allele frequency p for the underdominant case ($w_{11} = 1$, $w_{12} = 0.2$, $w_{22} = 0.8$). The jagged line shows how the allele frequency changes over successive generations as in Figure 6. In this example, the equilibrium at $\hat{p} = 0.428$ is unstable.

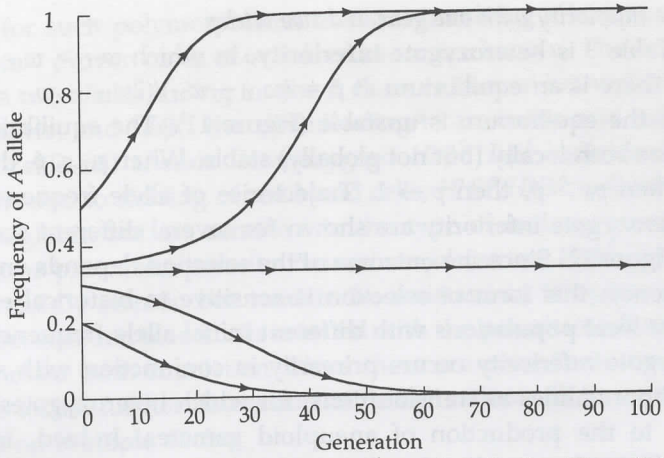


Figure 12. Selection when there is heterozygote inferiority. The allele frequency goes to 0 or 1, depending on the initial frequency. In this example, $w_{11} = 1$, $w_{12} = 0.8$, $w_{22} = 0.9$, and there is an unstable equilibrium when the frequency of the A allele is $p = 0.333$. An infinite population with $p = 0.333$ maintains this frequency, but any slight downward change in the frequency of A results in ultimate loss of A, while a slight increase in p over 0.333 results in fixation of A.

the rearrangement ought to become fixed automatically and carry along such genes as ones conferring sensitivity to an insecticide (see Figure 13). Specifically, the normal individuals may be looked upon as being homozygous for an "allele," say *N/N*, and the individuals with the rearrangement as being homozygous for an alternative "allele," say *R/R*. Because of the viability considerations mentioned above, the relative fitnesses of *N/N*, *R/N*, and *R/R* individuals may be taken as $1:0:1-s$, where s measures the fraction of inviable zygotes in $R/R \times R/R$ matings. This represents an extreme case of heterozygote inferiority with an unstable equilibrium at $\hat{p} = (1-s)/(2-s)$, where $\hat{p}$ represents the frequency of N among gametes. For $s = 3/4$ (a rather typical value for complex chromosomal rearrangements), $\hat{p} = 0.2$, or $\hat{q}$ (the frequency of R among gametes) is equal to 0.8. Thus, if an *N/N* population is flooded with *R/R*'s at a level of more than four *R/R*'s for every *N/N*, the *R/R* type would be expected to go to fixation. (For another application of complex chromosomal rearrangements in population genetics, see Jungen and Hartl 1979; Haymer and Hartl 1981, 1983b).

The Fundamental Theorem of Natural Selection

Fisher (1930) gave this somewhat pretentious name to an interesting property of one-locus viability selection models. In words, the theorem states that the rate of increase of mean fitness is equal to the genetic variance in fitness.

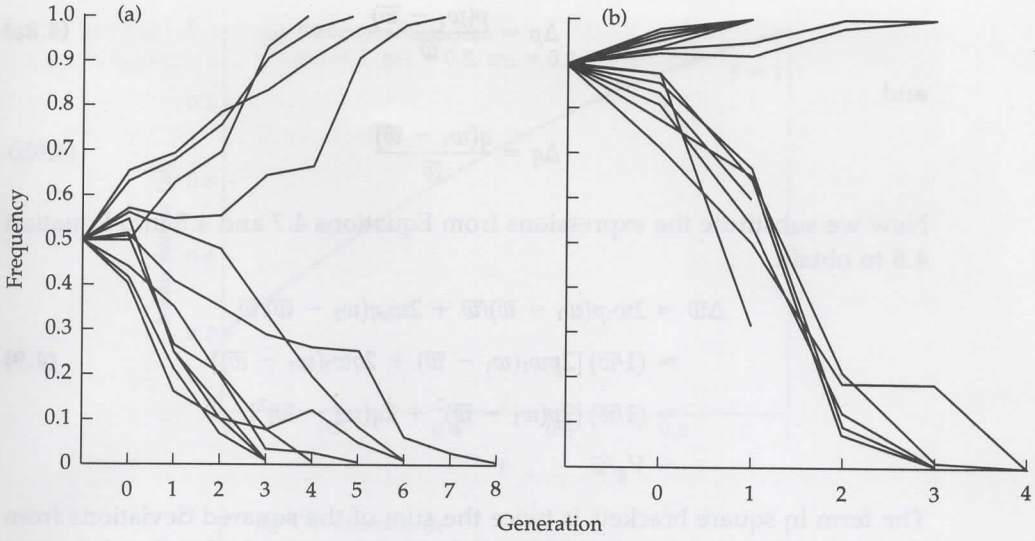


Figure 13. The change in chromosomal frequency over time when the heterozygote has a disadvantage and the unstable equilibrium is at 0.5 (a) or 0.9 (b) for translocations in *Drosophila melanogaster*. (From Foster et al. 1972.)

This theorem has several important consequences, including the implication that the mean fitness is nondecreasing, and that in order for there to be evolutionary change due to natural selection, there must be variation in fitness. We will follow a derivation of the fundamental theorem suggested by Ginzburg and Golenberg (1985). The method makes the approximation that $\Delta\overline{w} \approx d\overline{w}/dt$, which is reasonably accurate when selection is sufficiently weak that allele frequency does not change very rapidly. The formal proof of the fundamental theorem does not require this restriction. The first step requires an application of the chain rule in calculus, and is

$$\Delta\overline{w} \approx d\overline{w}/dt = \left(\frac{\partial\overline{w}}{\partial p}\right)\left(\frac{dp}{dt}\right) + \left(\frac{\partial\overline{w}}{\partial q}\right)\left(\frac{dq}{dt}\right) \tag{4.6}$$

This is just another way to write the time derivative of the mean fitness, or the rate of change of mean fitness. Note that

$$\begin{aligned}\frac{\partial\overline{w}}{\partial p} &= w_{11}\frac{dp^2}{dp} + w_{12}\frac{d(2pq)}{dp} + w_{22}\frac{dq^2}{dp} \\ &= 2(pw_{11} + qw_{12})\end{aligned} \tag{4.7}$$

The expression $(pw_{11} + qw_{12})$ represents an average effect of the A allele, and is called the **marginal fitness** of the A allele, denoted w_1. Equation 4.7 can be written $\partial\overline{w}/\partial p = 2w_1$, and similarly $\partial\overline{w}/\partial q = 2w_2$. Another way to write Equation 4.4 is $\Delta p = pq(w_1 - w_2)/\overline{w}$, which also implies that

$$\Delta p = \frac{p(w_1 - \overline{w})}{\overline{w}} \tag{4.8a}$$

and

$$\Delta q = \frac{q(w_2 - \overline{w})}{\overline{w}} \tag{4.8b}$$

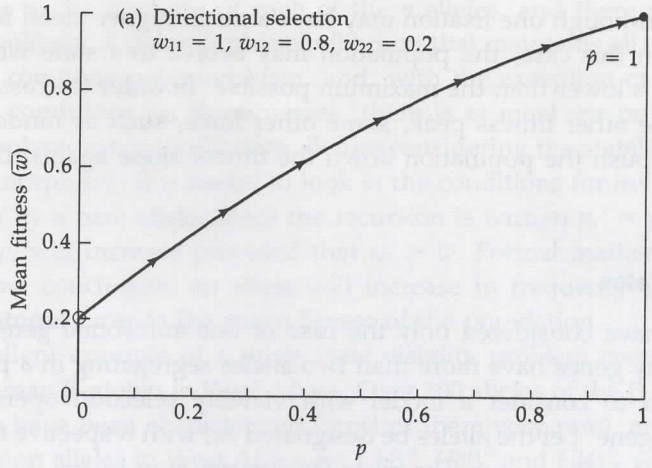
Now we substitute the expressions from Equations 4.7 and 4.8 into Equation 4.6 to obtain

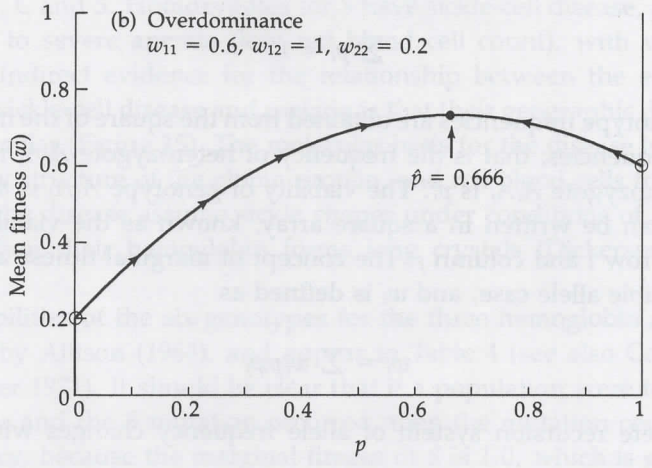
$$\begin{aligned} \Delta\overline{w} &\approx 2w_1p(w_1 - \overline{w})/\overline{w} + 2w_2q(w_2 - \overline{w})/\overline{w} \\ &\approx (1/\overline{w})\,[2pw_1(w_1 - \overline{w}) + 2qw_2(w_2 - \overline{w})] \\ &\approx (1/\overline{w})\,[2p(w_1 - \overline{w})^2 + 2q(w_2 - \overline{w})^2] \\ &\approx V_g/\overline{w} \end{aligned} \tag{4.9}$$

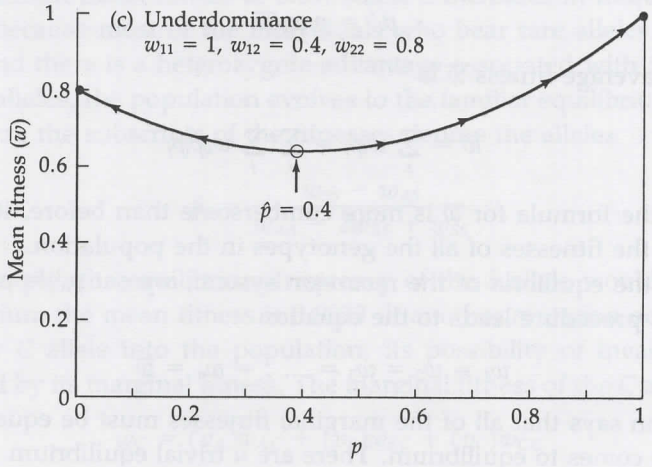
The term in square brackets is twice the sum of the squared deviations from the mean fitness weighted by the allele frequencies, which is by definition the genetic variance in fitness (V_g). The change in mean fitness is approximately equal to the genetic variance in fitness with appropriate scaling of fitnesses to make $\overline{w} = 1$. The approximation improves for smaller differences among the fitnesses, because with weaker selection, the approximation $\Delta\overline{w} \approx d\overline{w}/dt$ becomes more accurate. A helpful graphical way to view the fundamental theorem makes use of an **adaptive topography**, which is a surface of mean fitness plotted as a function of allele frequencies (Wright 1932, 1988). The fundamental theorem of natural selection states that, in viability models, the allele frequencies change in such a way that the population always goes "uphill" on the adaptive topography.

The fundamental theorem with two alleles can be illustrated by drawing plots of the mean fitness for all values of the allele frequency (Figure 14). In the case of directional selection, the mean fitness is maximized at the stable fixation, and the mean fitness constantly increases as the allele frequency approaches this fixation. With overdominance, the maximum in mean fitness occurs at precisely the stable equilibrium $\hat{p}$. With heterozygote inferiority, the minimum in mean fitness occurs at the unstable equilibrium, and since mean fitness always increases, if we start below the equilibrium, the allele frequency decreases, while if we start above it, the allele frequency increases to the fixation. Both fixations are local maxima in the case of heterozygote

Figure 14. Plots of mean fitness ($\overline{w}$) for the three possible patterns of viability selection at one locus with two alleles. (a) Directional selection, (b) overdominance, and (c) underdominance. The arrows indicate the direction of allele frequency change. ▷


(a) Directional selection
$w_{11} = 1, w_{12} = 0.8, w_{22} = 0.2$
$\hat{p} = 1$
Mean fitness ($\bar{w}$)
p


(b) Overdominance
$w_{11} = 0.6, w_{12} = 1, w_{22} = 0.2$
$\hat{p} = 0.666$
Mean fitness ($\bar{w}$)
p


(c) Underdominance
$w_{11} = 1, w_{12} = 0.4, w_{22} = 0.8$
$\hat{p} = 0.4$
Mean fitness ($\bar{w}$)
p

inferiority, although one fixation may have a much higher mean fitness than the other. In this case, the population may evolve to a state with a mean fitness that is lower than the maximum possible. In order to cross the valley to get to the other fitness peak, some other force, such as random genetic drift, must push the population down the fitness slope against the force of selection.

Multiple Alleles

So far we have considered only the case of one autosomal gene with two alleles. Many genes have more than two alleles segregating in a population, so we need to consider a model with viability selection operating on a multiallelic gene. Let the alleles be designated A_i, with respective frequencies p_i, where $i = 1, 2, \ldots, n$. The allele frequencies sum to 1:

$$\sum_i p_i = 1 \tag{4.10}$$

and the genotype frequencies are obtained from the square of the multinomial of allele frequencies; that is the frequency of heterozygote A_iA_j is $2p_ip_j$ and that of homozygote A_iA_i is p_i^2. The viability of genotype A_iA_j is w_{ij}, and the viabilities can be written in a square array, known as the viability matrix, with w_{ij} in row i and column j. The concept of marginal fitness also applies to the multiple allele case, and w_i is defined as

$$w_i = \sum_j w_{ij}p_ip_j \tag{4.11}$$

The complete recursion system of allele frequency changes with multiple alleles is

$$p_i' = p_iw_i/\overline{w} \tag{4.12}$$

where the average fitness $\overline{w}$ is

$$\overline{w} = \sum_i w_ip_i = \sum_i \sum_j w_{ij}p_ip_j \tag{4.13}$$

Although the formula for $\overline{w}$ is more cumbersome than before, it is still the average of the fitnesses of all the genotypes in the population.

To find the equilibria of the recursion system, we set $p_i' = p_i$ and solve for p_i. This procedure leads to the equation

$$w_1 = w_2 = w_3 = \ldots = w_n = \overline{w} \tag{4.14}$$

This relation says that all of the marginal fitnesses must be equal when the population comes to equilibrium. There are n trivial equilibrium points, cor-

responding to the fixations of each of the n alleles, and there may also be interior equilibria. A polymorphic equilibrium that maintains all of the alleles is called a **complete** polymorphism, and, with the exception of biologically unrealistic conditions on the fitnesses, there is at most *one* point that is a complete polymorphic equilibrium. Before considering the stability of a complete polymorphism, it is useful to look at the conditions for invasion of the population by a new allele. Since the recursion is written $p_i' = p(w_i/\overline{w})$, it is clear that p_i will increase provided that $w_i > \overline{w}$. Formal mathematics gives us the same conclusion: an allele will increase in frequency provided its marginal fitness exceeds the mean fitness of the population.

An excellent example of a three-allele viability problem comes from the data on human β-globin in West Africa. Over 100 alleles of the β-globin gene in humans have been characterized (most of them very rare), and the three most common alleles in West Africa are $Hb\beta^A$, $Hb\beta^C$ and $Hb\beta^S$, which will be denoted A, C and S. Homozygotes for S have sickle-cell disease, an affliction that leads to severe anemia (low red blood cell count), with very painful episodes. Indirect evidence for the relationship between the elevated frequency of sickle-cell disease and malaria is that their geographic distributions are very similar (Figure 15). The molecular basis for the disease is the abnormal tertiary structure of the globin protein, and red blood cells from individuals with the disease assume sickle shapes under conditions of low oxygen tension when their hemoglobin forms long crystals (Dickerson and Geis 1983).

The viabilities of the six genotypes for the three hemoglobin alleles were estimated by Allison (1964), and appear in Table 4 (see also Cavalli-Sforza and Bodmer 1971). It should be clear that if a population were to start with all A alleles and the S mutation occurred, then the mutation could increase in frequency, because the marginal fitness of S is 1.0, which is greater than the then current mean fitness of 0.89. Allele S increases in frequency when it is rare, because most of the individuals who bear rare alleles are heterozygotes, and there is a heterozygote advantage associated with S. With just these two alleles, the population evolves to the familiar equilibrium with two alleles where the subscripts of the fitnesses denote the alleles

$$\hat{p}_S = \frac{w_{SS} - w_{AS}}{w_{AA} - 2w_{AS} + w_{SS}} \qquad \textbf{(4.15)}$$

In this example, the equilibrium frequency of the S allele would be 0.1209. At equilibrium the mean fitness is 0.9033. If another mutation occurs, introducing the C allele into the population, its possibility of invasion is also determined by its marginal fitness. The marginal fitness of the C allele equals

$$w_C = (p_A)w_{AC} + (p_S)w_{SC} + (p_C)w_{CC} \qquad \textbf{(4.16)}$$

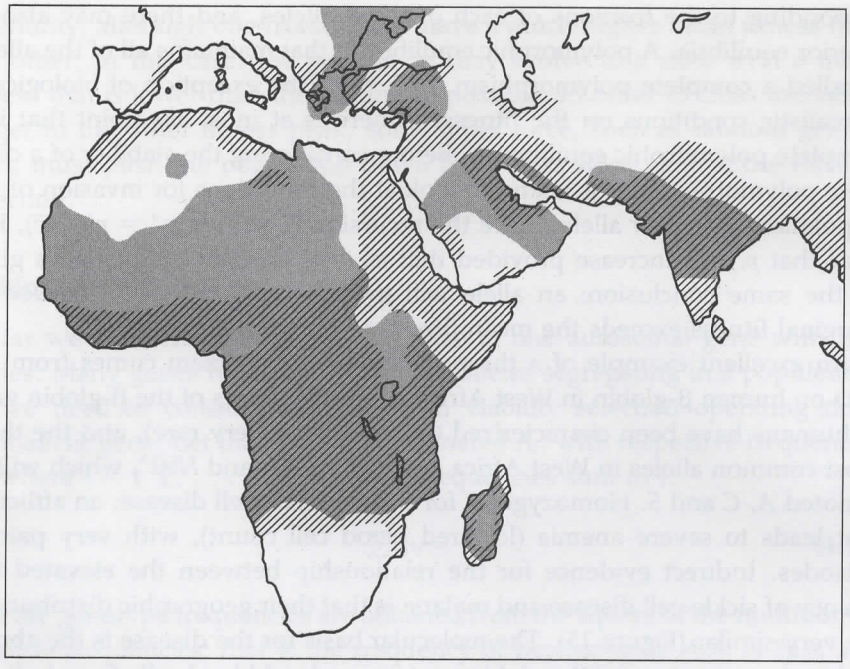


Figure 15. The hatched areas show the incidence of falciparum malaria around the 1920s (before control programs were implemented) in Africa, the Middle East, and southern Europe; the shaded regions designate areas of high incidence of sickle-cell hemoglobin disease. The extensive overlap in the distributions was an early indication that there might be some causal connection. (After Cavalli-Sforza 1974.)

When C is very rare the third term is very small because $p_C \approx 0$ and can be ignored, leaving $p_S = 0.1209$ and $p_A = 1 - p_S = 0.8791$. The marginal fitness of the C allele is therefore $(0.8791)(0.89) + (0.1209)(0.70) = 0.8670$. Since the marginal fitness of the C allele is less than the mean fitness, the C allele cannot invade when it is very rare. This example shows an important property of one-locus selection: even though fixation of the C allele would globally maximize the mean fitness, the population cannot achieve this state when C starts from a very low initial frequency.

Figures 16 and 17 present different ways to show the outcome of this three-allele problem graphically. In Figure 16, the instability of the A/S polymorphism is evident, and in Figure 17, note that if the C allele can be raised in frequency by a sufficient amount, its marginal fitness will exceed the mean fitness, and it can continue to increase in frequency and ultimately displace the A/S polymorphism. There is some room for skepticism in this prediction because the fitness estimates given in Table 4 are not without error, and they

Table 4. Observations of hemoglobin *A, S,* and *C* genotypes and Hardy-Weinberg expectations for 72 West African populations.

	GENOTYPE						
	AA	*SS*	*CC*	*AS*	*AC*	*SC*	**Total**
Observed count	25,374	67	108	5482	1737	130	32,898
Expected count	25,615.5	306.87	74.69	4967.2	1768.6	165.01	32,898
Fitness[a]	0.991	0.218	1.446	1.104	0.982	0.788	—
Standardized fitnesses and their standard errors[b]	0.89 ± 0.03	0.20 ± 0.11	1.31 ± 0.29	1	0.89 ± 0.035	0.70 ± 0.07	—

[a] Fitnesses were calculated as the ratio of the observed to the expected counts.
[b] Fitnesses standardized by taking $\overline{AS} = 1$.
(From Cavalli-Sforza and Bodmer 1971.)

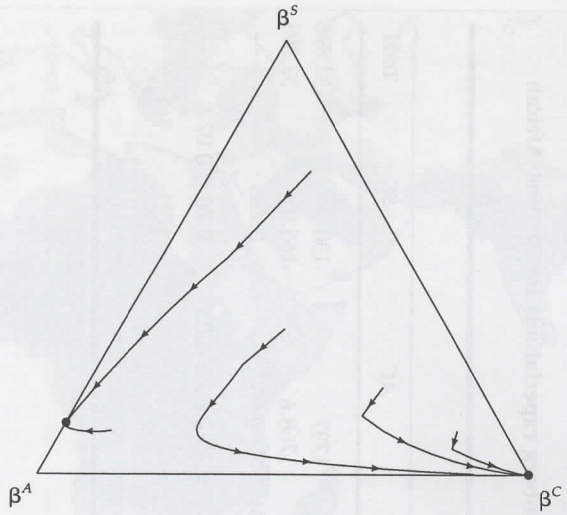


Figure 16. De Finetti diagram showing stability of β^C fixation and β^A–β^S polymorphism. Both of these equilibria are locally stable, but neither is globally stable.

may change when, for instance, the incidence of malaria changes. However, recent surveys show that the *C* allele does seem to be increasing in frequency and expanding its geographic distribution.

Resolving what happens when three alleles are segregating requires more mathematics. The equilibrium with all three alleles present must satisfy the equations

$$\begin{aligned} p_1' &= p_1(w_1/\overline{w}) \\ p_2' &= p_2(w_2/\overline{w}) \\ p_3' &= p_3(w_3/\overline{w}) \end{aligned} \tag{4.17}$$

so that at equilibrium $w_1 = \overline{w}$, $w_2 = \overline{w}$ and $w_3 = \overline{w}$. This is a system of three equations with three unknowns, and it has either no roots, a unique solution, or an infinity of solutions. An easy way to solve a system of n linear equations with n unknowns is to use Cramer's rule, which can be found in linear algebra textbooks. The conditions for there to be no solutions or an infinite number of solutions are biologically implausible, so in most cases there is a unique solution. However, this solution is not always biologically meaningful, because we require the additional constraint that $0 < p_i < 1$ for all allele frequencies. If the solution does not satisfy these conditions, then at least one of the alleles will be lost. That is, there is no valid complete polymorphism. If the complete polymorphic equilibrium exists and is biologically valid, then a result of Kingman (1961) provides a method for determining its stability. The condition for the stability of the complete polymorphism is that

the viability matrix **W** must have exactly one positive eigenvalue and at least one negative eigenvalue. If the complete polymorphism exists and is locally stable by Kingman's criterion, then its uniqueness guarantees global stability. This means that no matter what allele frequencies are used as a starting point (as long as they are not zero), the population will always evolve to the complete polymorphism. This conclusion also follows from the fundamental theorem of natural selection, since the mean fitness is nondecreasing.

The multiple allele viability model has some surprising properties. In the case of two alleles, a necessary and sufficient condition for a stable polymorphism was heterozygote advantage. Necessary and sufficient means that, if there is a stable polymorphism with two alleles, then the heterozygote must be the most fit; and if the heterozygote is the most fit, then there must be stable polymorphic equilibrium. In the case of three and more alleles, heterozygote advantage is more difficult to define because there are many more genotypes. One definition that seems reasonable is

$$w_{ii} < w_{ij} > w_{jj} \tag{4.18}$$

so that a heterozygote is more fit than the homozygote for either of its constituent alleles. However, this type of heterozygote advantage is *neither* necessary *nor* sufficient to guarantee a complete polymorphism. Nevertheless, the mean fitness at a stable equilibrium is greater than the fitness of any of the homozygotes.

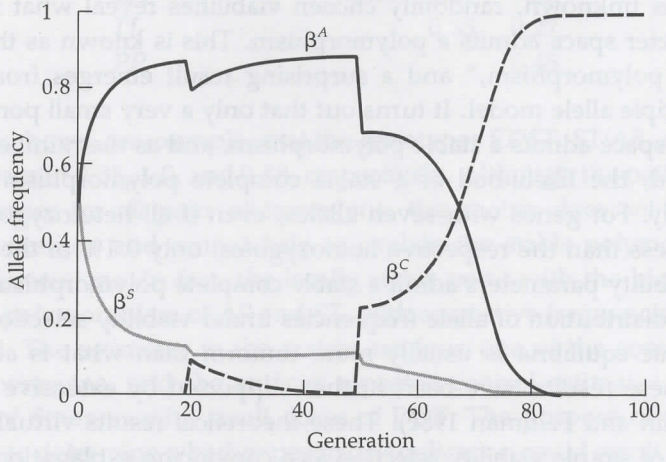


Figure 17. Hypothetical trajectories of β-globin allele frequencies, using the fitnesses given in Table 4. Initially the frequencies of alleles β^A and β^S are 0.5, and they come to a stable equilibrium. At generation 20 a small frequency of β^C is introduced, and it fails to invade. At generation 50, a higher frequency of β^C is introduced, and it increases in frequency until it is fixed.

How reasonable is it to assume that genetic variation is maintained in natural populations by processes such as the ones we have been modeling? During the 1970s a great wealth of data was collected on the amounts and geographic distributions of allozyme variation, and many explanations were sought for the observed variation. Since heterozygote advantage for two alleles is necessary and sufficient to guarantee a polymorphism, this suggested that overdominance might be an important mechanism for maintaining variation. However, the overdominance model was undercut by the demonstration that viability parameters resulting in the maintenance of multiple allele polymorphisms by heterozygote superiority are hard to find. This way of looking at multiallelic polymorphisms was explored by Lewontin et al. (1978) by deriving conditions for the stability of a complete polymorphism. The first set of conditions for stability is known as the "triangle inequality," which can be written

$$w_{ij} < w_{ik} + w_{jk} \tag{4.19}$$

These conditions, in combination with

$$w_{ij} > (w_{ii} + w_{jj})/2 \tag{4.20}$$

provide necessary conditions for a complete polymorphism. The stability conditions were rederived from basic principles by Mandel (1980). Lewontin et al. (1978) asked about the likelihood that multiple allele polymorphisms could be maintained by supposing that viabilities were chosen at random from the possible range of viabilities. Although the *a priori* distribution of viabilities is unknown, randomly chosen viabilities reveal what fraction of the parameter space admits a polymorphism. This is known as the "opportunity for polymorphism," and a surprising result emerges from the one locus–multiple allele model. It turns out that only a very small portion of the parameter space admits a stable polymorphism, and as the number of alleles is increased, the likelihood of a stable complete polymorphism decreases dramatically. For genes with seven alleles, even if all heterozygotes have a higher fitness than the respective homozygotes, only 0.1% of the randomly chosen viability parameters admit a stable complete polymorphism. Furthermore, the distribution of allele frequencies under viability selection at stable polymorphic equilibria is usually more uniform than what is actually observed. These results have been further supported by extensive numerical work (Karlin and Feldman 1981). These theoretical results virtually exclude the model of simple viability selection as a convincing explanation for maintaining multiple allele polymorphism.

Populations in nature are not simply presented with a single array of alleles; rather, mutations occur successively over long periods of time, and each new allele will increase in frequency under the influence of natural

selection if its marginal fitness exceeds the mean fitness. Stable multiple allele polymorphisms are much easier to find in numerical simulations that allow this kind of build-up of allelic diversity (Spencer and Marks 1988).

One classic experimental system that has been explained by the multiple allele viability selection model is the system of third chromosome inversions in *Drosophila pseudoobscura*, extensively studied by Theodosius Dobzhansky and his students. The most common third chromosome arrangement is called *Standard* (*ST*), and other arrangements are related to *Standard* by one or more inversions. Most populations are polymorphic for two and often for several inversion patterns. Dobzhansky and his co-workers estimated the fitnesses of the various karyotypes in a series of papers based on experiments using population cages. By allowing flies of the various karyotypes to compete in a pairwise manner, models were fitted whose parameters were the relative fitnesses of each karyotype. Large standard errors are associated with these estimates, and a number of statistical problems have been ascribed to this approach (Prout 1965); nevertheless, the method reveals strikingly large differences among the karyotypes. One such array of fitness estimates is given below for a series of third chromosome inversion patterns that are all found in the Mather, California population (Anderson et al. 1968).

	ST	*AR*	*CH*	*TL*	*PP*
ST	0.35	1.0	0.76	0.51	0.61
AR		0.58	0.26	0.45	0.33
CH			0.26	0.19	0.26
TL				0.35	0.27
PP					0.45

This matrix shows, for example, that the genotypes *ST/ST*, *ST/AR*, and *AR/AR* have fitnesses of 0.35, 1.0, and 0.58, respectively. Although the pattern shows overdominance for all pairs of inversions, this matrix does *not* satisfy the Kingman criterion, and cannot help to explain the stable polymorphism of these five inversions. In fact, the locally stable point with the highest mean fitness is a polymorphism of *AR* and *ST*, with respective frequencies of 0.6075 and 0.3925. The estimates in the matrix are from one of the control studies of these inversions, and the authors repeated control estimates as well as estimates of fitnesses with small doses of DDT. The purpose of the experiments was to determine whether pesticide application could result in changes in the frequencies of third chromosome inversions by changing relative fitnesses. The result was negative, because even though the fitness estimates appeared different, the control replicates were more different still, suggesting that the estimates had considerable uncertainty. The poor reliability of the

estimates was further evidenced by the large χ^2 values, which indicated a poor fit between the model and the actual frequencies in the population cages.

The Fundamental Theorem with Multiple Alleles

The fundamental theorem of natural selection has been proved in a generalized form applicable to an arbitrary number of alleles. Beginning with a numerical example, Figure 18 shows a plot of the mean fitness surface for the three globin alleles *A*, *S*, and *C*. Fitness is maximized when the *C* allele is fixed, and the region of minimum fitness defines a line of separation that natural selection alone cannot cross. We will not go through all of the steps, but sketch the outline of the algebraic proof, which is due to Kingman (1961). First, note that the mean fitness can be written

$$\bar{w} = \sum_i \sum_j w_{ij} p_i p_j \tag{4.21}$$

To obtain an expression for the change in mean fitness, consider first the changes in allele frequency, since genotype frequencies are obtained by random union of gametes and fitnesses are constant.

$$\bar{w}' = \sum_i \sum_j w_{ij} p_i' p_j' \tag{4.22}$$

Now recall that

$$p_i' = \frac{p_i \sum_k w_{ik} p_k}{\bar{w}} \tag{4.23}$$

With similar changing of subscripts, we can write p_j' as a function of the genotypic fitnesses and p_j. After substituting, we obtain

$$\begin{aligned} \bar{w}' &= \sum_i \sum_i w_{ij} \left(\frac{\sum_k w_{ik} p_k}{\bar{w}} \right) \left(\frac{\sum_l w_{jl} p_l}{\bar{w}} \right) p_i p_j \\ &= \bar{w}^{-2} \sum_i \sum_j \sum_k \sum_l w_{ij} w_{ik} w_{jl} p_i p_j p_k p_l \end{aligned} \tag{4.24}$$

This looks messy, and the mathematics gets worse before it gets better. However, Kingman (1961) was able to show from this relation that

$$\bar{w}' \geq \bar{w} \tag{4.25}$$

so the mean fitness is nondecreasing in the multiple allele case. With multiple

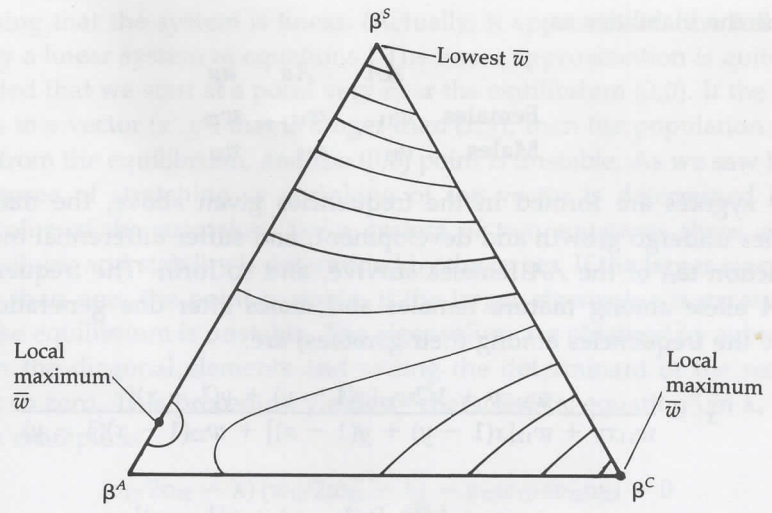


Figure 18. Mean fitness ($\overline{w}$) surface (also called an adaptive topography) for the examples given in Figures 16 and 17.

alleles the adaptive topography can have several local maxima. Because the mean fitness always increases, a population that is part of the way up a small hill cannot go down into a small valley in order to ascend to a higher peak, even if the other peak is much higher. Wright's "shifting balance theory," presented in Chapter 6, provides one mechanism by which populations can traverse fitness valleys on the way to higher peaks. Despite the Panglossian picture that the fundamental theorem of natural selection implies, many models violate the fundamental theorem, including the two-locus viability model, fecundity selection, and many types of ecological complications (Curtsinger 1984). Nevertheless, the fundamental theorem of natural selection has been influential.

Two-Sex Viability Selection

Suppose that the two sexes have different viabilities for the various genotypes. When this is the case we need to keep track of the allele frequencies in the two sexes separately. In particular, zygote frequencies are inferred from the products of frequencies of uniting gametes, assuming random union of gametes. Suppose that the frequencies of A and a in eggs are x and $1 - x$, respectively. Let y be the frequency of A in sperm and $1 - y$ be the frequency of a in sperm. The frequency of zygotes of both sexes is then

AA	*Aa*	*aa*
xy	$x(1 - y) + y(1 - x)$	$(1 - x)(1 - y)$

Define the viabilities as

	AA	*Aa*	*aa*
Females	w_{11}	w_{12}	w_{22}
Males	v_{11}	v_{12}	v_{22}

After zygotes are formed in the frequencies given above, the males and females undergo growth and development, and suffer differential mortality. A fraction w_{11} of the *AA* females survive, and so forth. The frequencies of the *A* allele among mature females and males after one generation (and hence the frequencies among their gametes) are:

$$x' = \frac{w_{11}xy + 1/2w_{12}[x(1-y) + y(1-x)]}{w_{11}xy + w_{12}[x(1-y) + y(1-x)] + w_{22}(1-x)(1-y)}$$

$$y' = \frac{v_{11}xy + 1/2v_{12}[x(1-y) + y(1-x)]}{v_{11}xy + v_{12}[x(1-y) + y(1-x)] + v_{22}(1-x)(1-y)} \qquad \textbf{(4.26)}$$

One could try to set $x' = x$ and $y' = y$ and solve for the equilibria, but this is not recommended because the equations cannot be solved directly. Instead, we will use a method that is very useful in analyzing complex models. The stability of the points $\hat{x} = 0$, $\hat{y} = 0$ and $\hat{x} = 1$, $\hat{y} = 1$ can be very informative, since instability of both of these fixation points implies that the population cannot go to fixation for either allele. Verifying the instability of these points, known as **protected polymorphism**, is almost as good as knowing about interior equilibrium stability. If a model has a protected polymorphism, then we have identified another means of maintaining genetic variation. From a practical point of view, the job of analyzing the model is done. Of course, by looking only at the fixation stabilities we may miss exciting aspects of the model, as there may be wild cycles and other bizarre dynamics.

To test the stability of the fixations, we **linearize** the system of equations. Linearizing is done by forming a matrix known as the Jacobian matrix, whose elements are the partial derivatives

$$\begin{bmatrix} \partial x'/\partial x & \partial x'/\partial y \\ \partial y'/\partial x & \partial y'/\partial y \end{bmatrix} \qquad \textbf{(4.27)}$$

These are pretty messy in general, but they simplify considerably when evaluated at the fixation (0,0). For the two-sex viability model, the Jacobian matrix at (0,0) is:

$$\begin{bmatrix} w_{12}/2w_{22} & w_{12}/2w_{22} \\ v_{12}/2v_{22} & v_{12}/2v_{22} \end{bmatrix} \qquad \textbf{(4.28)}$$

This matrix tells us how a vector (x,y) is transformed to a new vector (x',y')

assuming that the system is linear. (Actually, it approximates the full recursion by a linear system of equations.) The linear approximation is quite good provided that we start at a point very near the equilibrium (0,0). If the matrix results in a vector (x',y') that is longer than (x,y), then the population moves away from the equilibrium, and the (0,0) point is unstable. As we saw before, the degree of stretching or shrinking of the vector is determined by the eigenvalue of the recursion. For a system of two equations there are two eigenvalues, and stability is determined by the larger. If the larger eigenvalue is less than one, the point is stable; if the larger eigenvalue is greater than one, the equilibrium is unstable. The eigenvalues are obtained by subtracting λ from the diagonal elements and setting the determinant of the resulting matrix to zero. This procedure yields a "characteristic equation" in λ, which in this example is

$$(v_{12}/2v_{22} - \lambda)(w_{12}/2w_{22} - \lambda) - v_{12}w_{12}/4v_{22}w_{22} = 0 \qquad \textbf{(4.29)}$$

The roots of the characteristic equation are the eigenvalues, and in this case they are

$$\lambda = 0,\ \lambda = \frac{v_{12}w_{22} + v_{22}w_{12}}{2v_{22}w_{22}} \qquad \textbf{(4.30)}$$

The condition that the larger eigenvalue is larger than one is equivalent to

$$v_{12}/v_{22} + w_{12}/w_{22} > 2 \qquad \textbf{(4.31)}$$

When this procedure is repeated, solving the characteristic equation of the linearized system at the other fixation $\hat{x} = 1$, $\hat{y} = 1$, we find that the condition for instability of the other fixation is

$$w_{12}/w_{11} + v_{12}/v_{11} > 2 \qquad \textbf{(4.32)}$$

These conditions appear suspiciously similar to the simple condition for overdominance. However, if there is a sufficiently strong overdominance in one sex, then the other sex does not need to have overdominance. If we scale the fitnesses to $w_{12} = v_{12} = 1$, then the conditions for polymorphism are that the harmonic mean of w_{11} and v_{11} (which is $2/[1/w_{11}) + (1/v_{11})]$) and the harmonic mean of w_{22} and v_{22} be less than one. As Ewens (1979) pointed out, the harmonic mean is less than the arithmetic mean, so using the argument of opportunity for polymorphisms, the two-sex case has greater opportunity for polymorphism than the one-sex viability model. When viabilities were chosen randomly on a uniform distribution between 0 and 1, out of 10,000 sets of viabilities tested, 4,233 yielded a protected polymorphism. This is significantly higher than the expected count of random viabilities that would yield a polymorphism from a one-sex viability model (3,333). (One-third of the random viabilities in the one-sex model lead to polymor-

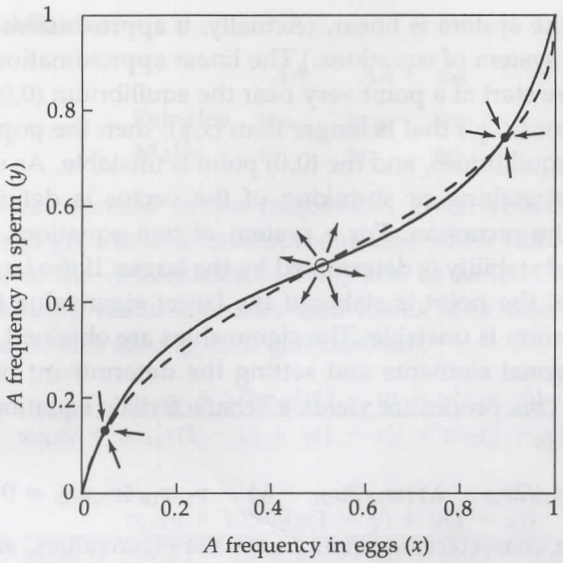


Figure 19. Phase plane showing the frequencies of the *A* allele in eggs (x) and in sperm (y) in the model with viability selection acting separately in the two sexes. The behavior of the model can be examined by setting $x' = x$ and $y' = y$ in Equation 4.26 and solving for two functions, $\hat{y} = f(x)$ (plotted as a solid line), and $\hat{x} = g(y)$ (dashed line). Points of intersection of these curves are equilibria of the system. In this example, the female viabilities of genotypes *AA*, *Aa*, and *aa* are $w_{11} = 1$, $w_{12} = 0.4$, $w_{22} = 0.9$, and the viabilities of the males are $v_{11} = 0.5$, $v_{12} = 1$, and $v_{22} = 0.6$ respectively. This case has two simultaneously stable equilibria, indicated by the solid dots, and one unstable equilibrium, indicated by the open circle. Note that there can be a stable equilibrium with underdominance in one sex, or opposite directional selection in the two sexes.

phism because, of the six possible orderings of the fitnesses w_{11}, w_{12}, and w_{22}, two will be overdominant.) The two-sex model is also of interest in allowing two simultaneously stable interior equilibria (Figure 19; further details can be found in Kidwell et al. 1977a).

X-Linked Viability Selection

For X-linked loci it is necessary to consider the effects on viability in both males and females. An X-linked locus with two alleles will allow three different female genotypes and two male genotypes, and we can assign viabilities

Females			Males	
AA	*Aa*	*aa*	*A*	*a*
w_1	w_2	w_3	v_1	v_3

In previous models we have assigned a value like p for the frequency of an allele, and recursions were worked out to yield the allele frequencies in successive generations. Many problems are somewhat simpler to solve, although sometimes harder to grasp initially, when ratios of allele frequencies are used. Haldane, one of the founders of theoretical population genetics, was particularly fond of this notation, and the X-linked model is one that is much simplified by following his approach (Haldane and Jayakar 1964). Consider the ratio of $A:a$ in eggs as u and the ratio of $A:a$ in sperm as v. The frequencies of the female zygotes AA, Aa, and aa are therefore in the ratio $uv:u+v:1$, and the male genotypes A and a have frequencies in the ratio $u:1$. By tabulating the gametes produced by all of the genotypes, we can derive the recurrence equations as

$$\begin{aligned} u' &= \frac{(u+v)w_2 + 2uvw_1}{(u+v)w_2 + 2w_3} \\ v' &= (v_1/v_3)u \end{aligned} \tag{4.33}$$

After setting $u' = u$ and $v' = v$, we do some algebra to obtain the equilibrium

$$\hat{u} = \frac{(v_1+v_3)w_2 - 2v_3w_3}{(v_1+v_3)w_2 - 2v_1w_1} \tag{4.34}$$

and $\hat{v}$ is obtained directly from the recursion. As Mandel (1959) showed, if both the numerator and denominator are positive, then the equilibrium is stable; while if both are negative the equilibrium can still be valid, but it is unstable. Converting back to alleles, let p be the frequency of the A allele in eggs, so that $u = p/(1-p)$. Then $p/(1-p) = \hat{u}$, and solving for p gives the equilibrium allele frequency

$$\hat{p} = \frac{v_3w_3 - \frac{1}{2}(v_1+v_3)w_2}{v_1w_1 + v_3w_3 - (v_1+v_3)w_2} \tag{4.35}$$

This provides a biologically more intuitive rule for the stability of the X-linked viability model. *A polymorphism is stable if the products of the viabilities of the two sexes show overdominance.* (To apply the rule, the average of the viabilities of the two male types replaces the nonexistent heterozygous male.) The condition for stability of the polymorphism is consequently

$$v_1w_1 < 1/2(v_1+v_3)w_2 > v_3w_3 \tag{4.36}$$

It should be noted that heterozygote advantage in females is neither necessary nor sufficient to guarantee stability of this equilibrium. Indeed, the opportunity for polymorphism, defined as the portion of the parameter space that will admit a stable polymorphism, is generally smaller for the X-linked case than for the autosomal case (Curtsinger 1980). This restriction is

reflected in a smaller variance in viability effects of X chromosomes extracted from *Drosophila* populations compared to extracted second chromosomes (Eanes et al. 1985). As in the case of X linkage in approaching Hardy-Weinberg equilibrium, this model shows an oscillatory approach to equilibrium. The expectations of the X-linked viability model, including the oscillatory approach to equilibrium, have been observed in experimental populations of *Drosophila* (Hedrick 1976). An example of the theoretical dynamics of the model is given in Figure 20. More details of the analysis of X-linked viability selection can be found in Bennett (1957) and Cannings (1967). An analog to the fundamental theorem of natural selection was also shown to hold for X-linked viability selection (Hartl 1972). In this case, however, it is the harmonic mean of the fitnesses of the two sexes that remains nondecreasing.

HAPLOID SELECTION

Selection in its simplest form occurs in haploid organisms such as bacteria. Consider two competing bacterial genotypes A and a with relative viabilities w_1 and w_2, respectively. Suppose that in generation t the frequencies of A and a are p_t and q_t, respectively, with $p_t + q_t = 1$. In the absence of selection, the frequencies remain the same through time, so we have

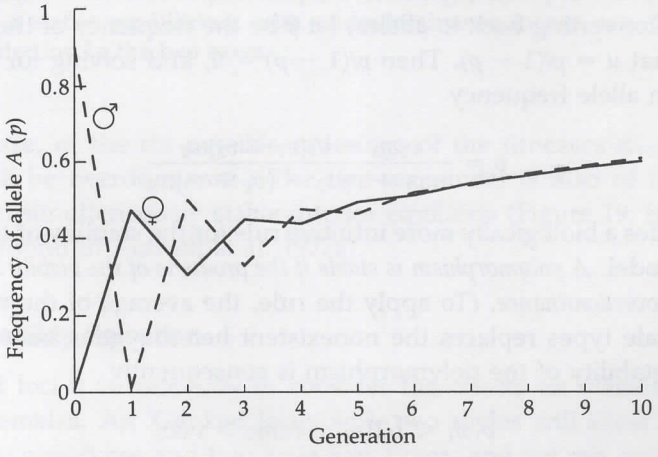


Figure 20. X-linked viability selection. The fitnesses in females are $w_1 = 0.8$, $w_2 = 1$, and $w_3 = 0.6$, and in males the fitnesses are $v_1 = 1$ and $v_3 = 0.9$. Note the oscillatory approach to equilibrium when the sexes have different initial allele frequencies. The allele frequencies continue to increase until they reach an equilibrium with $p_m = 0.7523$ and $p_f = 0.7321$, the latter being the theoretical value given by Equation 4.35.

$$p_t/q_t = p_t'/q_t' = p_0/q_0 \quad (4.37)$$

for any value of t. With selection, a proportion w_1 of the A genotypes survives, and a proportion w_2 of the a genotypes survives, so

$$p_1/q_1 = p_0w_1/q_0w_2 = (p_0/q_0)(w_1/w_2) \quad (4.38)$$

The relation between p_2/q_2 and p_1/q_1 is the same as that between p_1/q_1 and p_0/q_0, so $p_2/q_2 = (p_1/q_1)(w_1/w_2) = (p_0/q_0)(w_1/w_2)^2$. Continuing by generations in this manner, we have

$$p_t/q_t = (p_0/q_0)(w_1/w_2)^t \quad (4.39)$$

or

$$\ln(p_t/q_t) = \ln(p_0/q_0) + t\ln(w_1/w_2) \quad (4.40)$$

Box C and Figure 21 give an application of the equations for haploid selection. (For a more detailed analysis, see Kubitschek 1974.)

Observing Natural Selection in a Chemostat

C

By examining large populations for many generations, the most precise measure of selective differences among genotypes can be obtained. In a chemostat (a device for continuous culture of microbial populations, described in Chapter 2), *Escherichia coli* populations attain a steady state at between 10^6 and 10^9 cells/ml, depending on the flow rate and richness of the medium. Since it is fairly easy to follow a chemostat for 600 generations, this device has proven valuable for detecting very slight selective differences between pairs of genotypes. In practice, the two strains to be competed differ in resistance to the bacteriophage T5. Resistance to T5 is neutral under almost all conditions. Samples are taken from the chemostat and plated onto both nonselective plates and onto T5 bearing plates. The former give the total cell density, while the latter indicate the density of T5 resistance cells, and the ratio gives the desired gene frequency.

Dykhuizen and Hartl (1980) constructed a series of stocks having six naturally occurring alleles of the *gnd* locus (which encodes the enzyme 6PGD) into a common genetic background. In chemostats with gluconate-limited medium, four of these alleles showed no difference in growth rate, even though a difference as small as 0.005 hr^{-1} could be detected. One allele was detrimental to growth in gluconate medium but neutral when glucose was present. These results underscore the complexity of natural selection in diverse habitats (Hartl and Dykhuizen 1981). Similar experiments have been done on allelic variants that cause allozyme polymorphisms in PGI and G6PD (Dykhuizen and Hartl 1983; Dykhuizen et al. 1984a), and in the case of G6PD, no selective differences could be seen even though a growth rate difference of 0.002 hr^{-1} would have been detected.

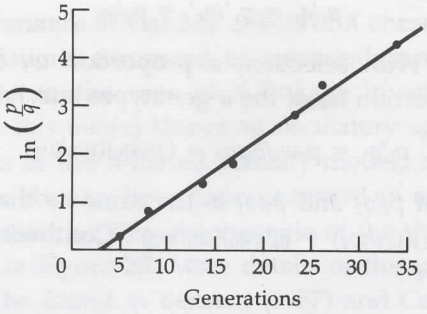


Figure 21. An example of the linear change in the natural log of p/q in a bacterial chemostat.

FREQUENCY-DEPENDENT SELECTION

Thus far we have considered fitness to be an invariant property of genotypes. If an individual had genotype AA, then its fitness was w_{11}, regardless of any other conditions in the population. In reality, a genotype may have different fitnesses depending on many aspects of the environment, including food or light availability, population density, and the relative frequencies of other genotypes (frequency-dependent selection). Perhaps the best example of frequency-dependent fitness is gametophytic self-incompatibility systems in many plants. This form of self-incompatibility is often due to one gene, with several alleles. Pollen of genotype S_i cannot successfully fertilize any female plant that bears the S_i allele. Any rare genotype will have the greatest number of potential successful mates, the fewest number of aborted pollinations, and therefore the highest fitness. Another example of frequency-dependent selection is mating success in *Drosophila*, where rare males sometimes seem to be at an advantage (Ehrman 1970; Spiess 1968, 1970; Petit and Ehrman 1969; Anderson 1969; Anderson and Brown 1984).

Perhaps the simplest scheme of frequency dependence is to make the three genotypes of one diallelic locus be linear functions of genotype frequency. In this particular scheme, the fitnesses would be

AA	Aa	aa
$1 - tp^2$	$1 - 2tpq$	$1 - tq^2$

With this pattern of frequency dependence, as any genotype frequency approaches 0, its fitness approaches 1; and if a genotype were to increase in frequency, its fitness would decrease (Figure 22). The equilibrium has a form that should be familiar by now (Lewontin 1958):

$$\hat{p} = \frac{w_{22}(p) - w_{12}(p)}{w_{11}(p) - 2w_{12}(p) + w_{22}(p)} \qquad \textbf{(4.41)}$$

In this case, the point $\hat{p} = \hat{q} = 1/2$ is a globally stable equilibrium. One interesting feature of this equilibrium is that the fitnesses are $w_{11} = w_{22} = 1 - (1/4)t$ and $w_{12} = 1 - (1/2)t$, which is an example of a stable interior equilibrium with a heterozygote disadvantage! With appropriate weighting, an analog to the fundamental theorem is valid under frequency-dependent viabilities (Curtsinger 1984).

Frequency dependence of viabilities has been observed in a number of experimental tests with *Drosophila* (Anderson et al. 1986; Ayala and Campbell 1974; Dawood and Strickberger 1969; Lewontin 1955; Nunney 1983; Snyder and Ayala 1979). The interpretation of these results is that differential utilization of resources by different genotypes, and larval competition for a limiting resource, sometimes result in frequency-dependent viability. A particular case with a fairly simple mechanism was examined by Hickey (1984). The absolute viability of *Amylase*-null larvae on starch medium was improved by the presence of larvae with active α-amylase, because the latter excreted enough active α-amylase to hydrolyze the starch for the larvae that lacked α-amylase activity.

The constant viability model fails to provide a plausible mechanism for maintaining more than a few alleles in a population because the opportunity for multiple allele polymorphism is small. By considering fitness as a property of genotype frequency, and in particular by allowing rare genotypes to have higher fitness, the opportunity for polymorphism might be increased somewhat. However, it remains a difficult experimental problem to determine the biological relevance of frequency dependence.

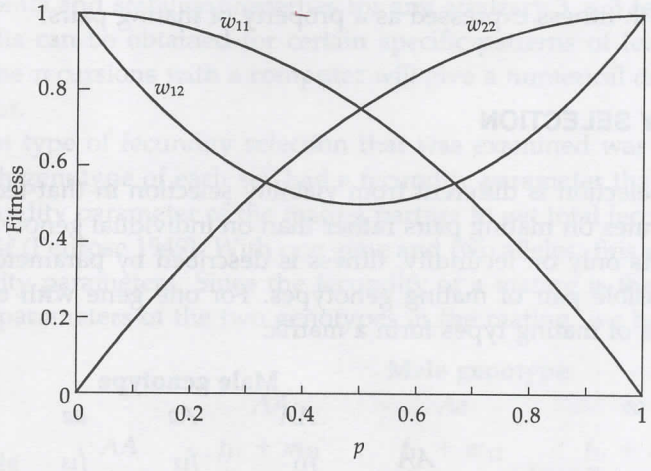


Figure 22. The model of frequency-dependent selection allows three genotypes to have fitnesses that vary with frequency, as plotted in the graph. Note that when the *A* allele is rare, the *AA* genotype has the highest fitness. In this example, the fitnesses are $w_{11} = 1 - p^2$, $w_{12} = 1 - 2pq$, and $w_{22} = 1 - q^2$.

Even the most widely cited example of frequency dependence — rare male mating advantage in *Drosophila* — has produced equivocal results. If rare males have an advantage, then every male should have some sort of advantage because every genotype is unique in *some* way. Markow (1978) tested the mating advantage of isogenic lines that differed in the region immediately flanking the *sepia* gene, with the rest of the genome identical. Even in this case, there was a rare male advantage. However, the rare male advantage is not found in all strains (Markow et al. 1980). Whether it is actually the *genotype* that is being assayed in the determination of rareness was tested by Bryant et al. (1980), who found that experimental clipping of a male's wings could introduce a rare male advantage. In short, the phenomenon is not an easy one to assay; there is much opportunity for experimental bias, and we cannot yet claim the general significance for rare male advantage.

Many situations that might appear to involve frequency-dependent selection can be more easily modeled with constant parameters. Incompatibility of Rh blood groups in humans can result in fetal loss, because Rh^- females who have been exposed to Rh^+ blood (by transfusion or previous pregnancy) produce anti-Rh antibodies that cross the placental barrier. This might appear to be a case of frequency dependence, because the fitness of an Rh^- female would depend on the frequencies of available mates. A frequency-dependent model could be devised for this situation, but it is more straightforward to develop the model in terms of the fecundities of the different matings (Feldman et al. 1969). Other frequently cited situations that suggest frequency-dependent selection, including self-incompatability in plants and selection on major histocompatibility loci (Hedrick and Thomson 1988), can also be modeled with fitness expressed as a property of mating pairs.

FECUNDITY SELECTION

Fecundity selection is different from viability selection in that fecundity selection operates on mating pairs rather than on individual genotypes. When selection acts only on fecundity, fitness is described by parameters relating to each possible pair of mating genotypes. For one gene with two alleles, the fitnesses of mating types form a matrix:

		Male genotype		
		AA	*Aa*	*aa*
Female genotype	*AA*	f_{11}	f_{12}	f_{13}
	Aa	f_{21}	f_{22}	f_{23}
	aa	f_{31}	f_{32}	f_{33}

(4.42)

Table 5. Mating table for fecundity selection.

Mating	Fecundity	Frequency	PROGENY		
			AA	*Aa*	*aa*
$AA \times AA$	f_{11}	x^2	1	0	0
$AA \times Aa$	f_{12}	xy	1/2	1/2	0
$AA \times aa$	f_{13}	xz	0	1	0
$Aa \times AA$	f_{21}	xy	1/2	1/2	0
$Aa \times Aa$	f_{22}	y^2	1/4	1/2	1/4
$Aa \times aa$	f_{23}	yz	0	1/2	1/2
$aa \times AA$	f_{31}	xz	0	1	0
$aa \times Aa$	f_{32}	yz	0	1/2	1/2
$aa \times aa$	f_{33}	z^2	0	0	1

By tabulating the mating types and the genotypes of the offspring produced by each mating, it is possible to construct the recursions for fecundity selection (Table 5). These recursions are in terms of genotype frequencies rather than allele frequencies, because the genotype frequencies cannot be inferred from the allele frequencies. In the case of viability selection, zygotes are formed each generation in Hardy-Weinberg proportions because of random mating, but fecundity selection destroys this property. The fact that fecundity selection must be modeled in terms of genotype frequencies makes it algebraically complex, and the one-gene two-allele fecundity model has not been solved in general. This means that we cannot simply write down the equilibrium points and stability properties for any arbitrary 3×3 fecundity matrix. Results can be obtained for certain specific patterns of fecundity, and iterating the recursions with a computer will give a numerical description of its behavior.

The first type of fecundity selection that was examined was additive, in which each genotype of each sex had a fecundity parameter that was added to the fecundity parameter of the mating partner to get total fecundity of the mating pair (Penrose 1949). With one gene and two alleles, this would imply six fecundity parameters. Since the fecundity of a mating is the sum of the fecundity parameters of the two genotypes in the mating, we have

		Male genotype		
		AA	*Aa*	*aa*
Female genotype	*AA*	$f_{11} + m_{11}$	$f_{11} + m_{12}$	$f_{11} + m_{22}$
	Aa	$f_{12} + m_{11}$	$f_{12} + m_{12}$	$f_{12} + m_{22}$
	aa	$f_{22} + m_{11}$	$f_{22} + m_{12}$	$f_{22} + m_{22}$

(4.43)

For the additive fecundity model the recurrence equations are:

$$\begin{aligned}
\bar{f}x' &= x^2(f_{11} + m_{11}) + 1/2(f_{11} + m_{12} + f_{12} + m_{11})xy + (1/4)y^2(f_{12} + m_{12}) \\
&= [(f_{11} + m_{11})x + 1/2(f_{12} + m_{12})y][x + (1/2)y] \\
\bar{f}y' &= 1/2(f_{11} + m_{12} + f_{12} + m_{11})xy + (f_{11} + m_{11} + f_{22} + m_{11})xz + 1/2(f_{12} + m_{12})y^2 \\
&\quad + 1/2(f_{12} + m_{22} + f_{22} + m_{12})yz \\
&= [(f_{11} + m_{11})x + 1/2(f_{12} + m_{12})y][z + (1/2)y) + [1/2(f_{12} + m_{12}) \\
&\quad + (f_{22} + m_{22})z](x + 1/2y) \qquad (4.44) \\
\bar{f}z' &= 1/4y^2(f_{12} + m_{12}) + 1/2(f_{12} + m_{22} + f_{22} + m_{12})yz + z^2(f_{22} + m_{22}) \\
&= [1/2y(f_{12} + m_{12}) + (f_{22} + m_{22})z](1/2y + z)
\end{aligned}$$

where $\bar{f}$ is the mean fecundity, calculated by summing the right hand sides of Equations 4.44. In the second set of recursions, the equations have been simplified to factor out terms like $x + 1/2y$. These are the allele frequencies, and can be envisaged as the allele frequencies in gametes produced by one sex. The other factor in the recursions is also like an allele frequency, but with weights applied to the genotype frequencies. These are precisely analogous to the situation where viability selection occurs in only one sex. It happens that the necessary and sufficient conditions to maintain a polymorphism in the one-sex situation is heterozygote advantage. Thus, an analogous principle holds for additive fecundity selection.

This method of recasting a problem that is new in terms of a problem to which the solution is known is a powerful tool. Since the behavior of a number of complex models is known, it is quite often the case that new problems can be expressed in terms of another model that has been solved.

Another pattern of fecundity selection that has been considered is the multiplicative model, in which the fecundity of a mating is the product of the fecundity effects in each sex (Bodmer 1965). The fecundity matrix in this case looks like

		Male genotype		
		AA	*Aa*	*aa*
Female genotype	*AA*	$f_{11}m_{11}$	$f_{11}m_{12}$	$f_{11}m_{22}$
	Aa	$f_{12}m_{11}$	$f_{12}m_{12}$	$f_{12}m_{22}$
	aa	$f_{22}m_{11}$	$f_{22}m_{12}$	$f_{22}m_{22}$

(4.45)

and the recursion system is obtained from the mating table as before. After some algebra, the recursion system becomes

$$\begin{aligned}
\bar{f}x' &= (xf_{11} + 1/2f_{12}y)(xm_{11} + 1/2m_{12}y) \\
\bar{f}y' &= (xf_{11} + 1/2f_{12}y)(1/2m_{12}y + m_{22}z) + (1/2f_{12}y + f_{22}z)(xm_{11} + 1/2m_{12}y) \qquad (4.46) \\
\bar{f}z' &= (1/2f_{12}y + f_{22}z)(1/2m_{12}y + m_{22}z)
\end{aligned}$$

This recursion also factors in a way that makes clear the similarity with viability selection. The frequencies of the genotypes are determined each generation by the product of the gamete frequencies produced by males and females, but the adult genotype frequencies are weighted by the fecundities. This situation is analogous to the two-sex viability model. The only difference is that the population is typically examined at the gamete stage in viability models, and the recursion expressed in terms of allele frequencies, while in fecundity models the population is censused at the stage of mature adults. Nevertheless, the equilibrium structure of the multiplicative fecundity model is the same as the two-sex viability model, with the implication that there can be up to three interior equilibria with two simultaneously stable.

The third pattern of fecundities is the symmetric case, which was analyzed in detail by Hadeler and Liberman (1975). In this model the productivity of a mating depends on the degree of heterozygosity of mating pairs, as follows:

		Male genotype		
		AA	*Aa*	*aa*
Female genotype	*AA*	α	β	γ
	Aa	β	δ	β
	aa	γ	β	α

(4.47)

Hadeler and Liberman (1975) demonstrated that this model allows five polymorphic equilibria — three symmetric equilibria with equal numbers of the two homozygotes, and two asymmetric equilibria (Figure 23). The two fixa-

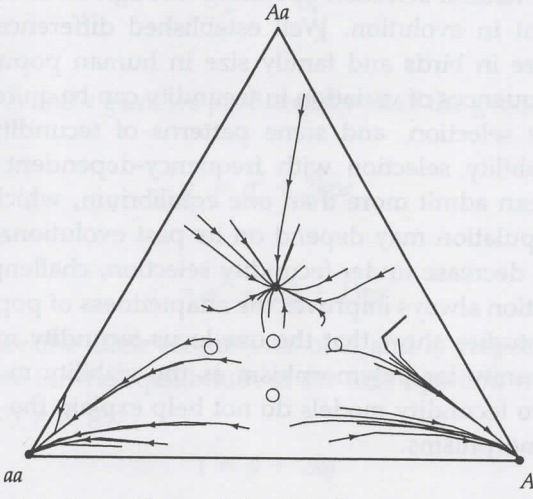


Figure 23. Symmetric fecundity model with five internal equilibria $\alpha = 16$, $\beta = 0.75$, $\gamma = 1$, and $\delta = 63$. In this particular example, which is not very biologically reasonable, the two fixations and one internal equilibrium are all stable, and there are four unstable equilibria, indicated by open circles. (From Hadeler and Liberman 1975.)

tions and all interior equilibria can be simultaneously stable, and the fate of the population in this case depends on the initial conditions. Feldman et al. (1983) found that the two asymmetric equilibria can also be simultaneously stable, and demonstrated conditions in which the symmetric fecundity model is equivalent to the multiplicative fecundity model.

Several points should be noted about the general fecundity model. Because mean fecundity can decrease over generations, there is no equivalent to the fundamental theorem of natural selection (Pollak 1978). A decrease in mean fecundity is shown in the following example. Consider the fecundity matrix

		Male genotype		
		AA	*Aa*	*aa*
Female genotype	*AA*	0	4	0
	Aa	4	8	0
	aa	0	0	0

(4.48)

If the population were started at Hardy-Weinberg genotypic proportions with $p = 0.9$, then the mean fecundity the next generation would be 3.818. Over the next two generations, the mean fecundity would be 3.476, then 3.425. The mean fecundity continues to decrease monotonically all the way to its equilibrium value of 3.414, at which point the equilibrium allele frequency is $\hat{p} = 0.6036$. Although random fecundity matrices will generally yield increasing mean fecundity, exceptions like this are not hard to find, and they prove an important shortcoming of the fundamental theorem of natural selection.

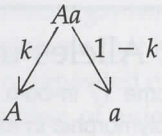
The process of natural selection operating through differential fecundities may be important in evolution. Well established differences in fecundity include clutch size in birds and family size in human populations. As we have seen, consequences of variation in fecundity can be quite different from those of viability selection, and some patterns of fecundity selection are equivalent to viability selection with frequency-dependent viabilities. Fecundity models can admit more than one equilibrium, which suggests that the state of a population may depend on its past evolutionary history. The mean fitness can decrease under fecundity selection, challenging the notion that natural selection always improves the adaptedness of populations. However, numerical studies show that the one-locus fecundity model has about the same opportunity for polymorphism as the viability model (Clark and Feldman 1986), so fecundity models do not help explain the maintenance of multiallelic polymorphisms.

MEIOTIC DRIVE AND GAMETIC SELECTION

Diploid organisms generally have a haploid phase in the life cycle, and selection operating in the haploid phase has markedly different properties

from selection in the diploid phase. Two forms of selection that occur in the haploid phase include meiotic drive, defined as the non-Mendelian production of gametes by heterozygotes, and gametic selection, which in plants is usually due to differential viability of gametes. To the uninitiated, meiotic drive might sound like blasphemy, but a number of examples occur in natural populations. The two most famous examples are the *t* alleles in mice (Lewontin and Dunn 1960; Young 1967; Silver 1985; see Box D) and the *SD* system in *Drosophila melanogaster*, both of which have been shown to be genetically complex (Hartl and Hiraizumi 1976; Hartl 1980; Brittnacher and Ganetzky 1984; Temin and Marthas 1984). Part of the complexity stems from the existence of autosomal and X-chromosomal modifiers that suppress the segregation effects of *SD* (Hiraizumi and Thomas 1984). The population genetic theory of such modifier genes was examined by Prout et al. (1973).

To model meiotic drive, consider one gene with two alleles, with heterozygotes producing a fraction k of A alleles and $1 - k$ of a alleles (so that $k = 1/2$ represents Mendelian segregation). The model assumes no difference between the sexes in the segregation parameter. In practice, meiotic drive usually occurs in only one sex — the male in the examples of the *t* alleles and *SD*. Such sexual asymmetry can easily be incorporated into the following model, but for now assume that heterozygotes of both sexes produce gametes as follows:



(4.49)

Considering the A and a gametes produced by all of the genotypes, we obtain the recursions

$$p' = p^2 + 2kpq$$

and

$$q' = q^2 + 2(1 - k)pq \qquad \textbf{(4.50)}$$

where p and q are the allele frequencies of A and a, respectively, and $p' + q' = 1$. If we solve for the equilibrium of the first equation we set $p' = p$ and divide through by p to get

$$1 = \hat{p} + 2k\hat{q} \qquad \textbf{(4.51)}$$

Since $q = 1 - p$, this equation implies that

$$1 = \hat{p} + 2k(1 - \hat{p}) = \hat{p}(1 - 2k) + 2k \qquad \textbf{(4.52)}$$

so $\hat{p} = 1$. The equation for p' also allows $\hat{p} = 0$ as a solution, so it appears

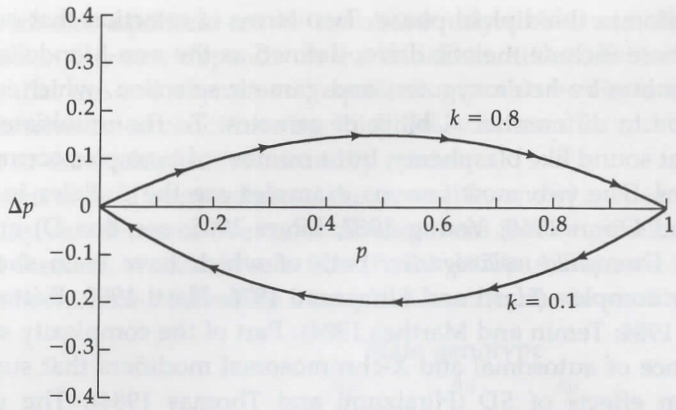


Figure 24. Plot of Δp versus p for the model of meiotic drive. The absence of an internal equilibrium is clear from the figure. If $k > 1/2$ then the A allele increases to fixation, while if $k < 1/2$, the A allele is lost.

that the only equilibria are the fixations. You should try to convince yourself that if $k > 1/2$, the A allele goes to fixation, whereas if $k < 1/2$, the a allele goes to fixation. In this version of meiotic drive, one allele or the other always goes to fixation (Figure 24).

D Meitoic Drive of the *t* Alleles in *Mus*

Near the centromere of chromosome 17 in both *Mus musculus* and *Mus domesticus* there is a simple *T* locus that is polymorphic in many populations for alleles that cause taillessness. The *T* locus is embedded in the *t* complex, a region of chromosome 17 that includes the major histocompatibility complex and many other genes; the *t* complex totals about 1% of the mouse genome (Silver 1985). Often associated with the *t* mutations are pleiotropic effects involving fertility, embryogenesis, meiotic recombination, and (most noticeably) male segregation ratio. The genes of the *t* complex are fairly tightly linked, so that we refer to the transmission of the group of genes as a "*t* haplotype." Males carrying a "complete" *t* haplotype in heterozygous form with a wild haplotype will transmit the *t* haplotype to over 90% of their offspring. The mechanism for the meiotic drive is still not clear, since *t*/+ males produce both sperm types in Mendelian proportions, and both appear normal. Not all *t*-bearing males show distorted segregation in their sperm, and molecular evidence shows that those showing Mendelian segregation bear "partial" *t* haplotypes. Such partial *t* haplotypes are exceedingly rare in nature, because there is strong viability selection against them, and they have no meiotic advantage. Molecular evidence is accumulating that suggests an analogy to the *Drosophila* "segregation distorter" system, which exhibits meiotic drive due to interactions between distorter and responder elements, both of which must be on the same chromosome in order to manifest drive (Silver and Remis 1987). Mice that are homozygous

for two complete *t* haplotypes are sterile, but again, the physiological mechanism is not entirely clear. About 73% of *t* haplotypes carry a recessive lethal, so many but not all *t/t* genotypes are lethal. The suppression of recombination appears to be due to a pair of inversions within the *t* complex, and recombination appears normal between nonlethal (complementing) *t/t* genotypes (Frischauf 1985).

It would appear that the meiotic drive favoring *t* alleles might be balanced in a population by the sterility or lethality of the *t/t* genotype. Lewontin and Dunn (1960) attempted this analysis long before the detailed genetic data on this complex genetic region were available. They found an average segregation parameter of $k = 0.952$, and obtained an expected equilibrium allele frequency for lethal *t* alleles of 0.385. Most natural populations have a much lower *t* frequency than this, suggesting that the model was overlooking something. We now see that the complex, multilocus genetic basis of *t* haplotypes accounts for some of the discrepancy. The stability of the polymorphism is indicated by the discovery (using molecular analysis) of very similar *t* haplotypes in several species of *Mus*, suggesting that the polymorphism predated the speciation (Delarbre et al. 1988).

Because the meiotic drive model predicts rapid fixation of the driven allele, the maintenance of polymorphism in natural populations must be through additional mechanisms. Possible means of maintaining meiotic drive alleles in populations include opposing selection (discussed later in this chapter) and polymorphism for other genes that modify the meiotic drive.

Gametic selection in monogamous organisms is modeled in the same way as meiotic drive. However, in organisms that have multiple mates, including plants that have mixtures of pollen deposited on the stigma, gametic selection follows a different model. In the case of animals, gametic selection is generally referred to as sperm competition, and in plants it is called pollen tube competition. The most striking forms of selection affecting the pollen tube are incompatibility mechanisms, which are not strictly the same as gametic selection because they typically involve interactions between maternal genotype and pollen haplotype.

Sperm displacement is another form of selection that can occur in polygamous organisms, and it is distinct from both meiotic drive and gametic selection. Sperm displacement refers to the degree to which the last male to mate with a female fathers the most offspring. In experiments with three genetically marked strains of *Drosophila melanogaster*, Prout and Bundgaard (1977) showed that the degree of sperm displacement differed markedly among strains. A male with greater sperm displacement is clearly going to be at a selective advantage. Fitness in this scenario is a property of the male genotype, not of the haploid sperm, and this is why sperm displacement has different population genetic properties. In a simple model of the process, Prout and Bundgaard (1977) showed that a one-locus two-allele polymor-

phism can be maintained if either (1) the heterozygote has the greatest sperm displacing ability, or (2) there are certain nontransitive conditions on displacing ability ($AA > Aa$, $Aa > aa$, and $aa > AA$).

SEXUAL SELECTION

In his book *The Descent of Man and Selection in Relation to Sex*, Darwin (1871) described two means by which selection could operate through differential mating. The first mechanism came from observations of male–male contests for mates. If genes that either confer greater ability to secure females or prevent other males from mating are introduced into a population, these characteristics could result in an increased frequency of alleles. The contests between males might be actual fights that inflict bodily harm to the opponent, or they might be primarily display or threat. Bighorn sheep males use several strategies in their quest to secure mates. The most conspicuous are the head-butting contests, which occur with impressive force but rarely result in injury. The winner of these contests gets access to females. Another strategy of male bighorns is "blocking," whereby the males try to herd females to the outer edge of their territory, thereby avoiding the headache of the butting contests while keeping the females to themselves (Hogg 1984). Darwin argued that male contests were the likely form of sexual selection in situations where females were a limiting resource, since males are then thrust into a situation where they must compete with other males for mates.

Distinct from the above example of male–male competition is a form of sexual selection that occurs when one sex chooses its mate. Darwin argued that mate choice is expected to occur in the sex that invests more energy in reproduction, which in most animals is the female. Because the female exerts a sizable effort in reproduction, it would be advantageous to her if she could choose a mate so as to maximize the fitness of her offspring. If reproduction costs males little time and effort, this line of reasoning suggests that they need not be so choosy. This form of sexual selection was inferred from the brilliant plumage in the males of many bird species. If the bright plumage was initially associated with good nutrition, or perhaps better genetic endowment for survival, then natural selection would augment the choosiness of the females. If the tendency for females to choose particular types of males is heritable, then this characteristic would increase in frequency because the sons of choosy females would carry both the traits that are desired in males *and* genes for female choosiness. This scenario results in Fisher's "runaway" sexual selection, a process proposed by Fisher (1958) in which both the chosen character and the choosing tendency increase to fixation. O'Donald (1978, 1979) has carried out computer simulations of this process and dem-

onstrated the critical dependence of linkage disequilibrium between the gene determining female choice and the gene determining a male character. An interesting feature is that the trait that is chosen in males can have a pleiotropic effect that is deleterious to viability, but the character can still increase in frequency. Another feature is that sons of chosen males are more likely to possess the desirable trait. Since males with desirable traits will be chosen in more matings, the number of genes passed to the next generation can exceed that of the alternative allele even if the matings result in fewer offspring per mating. Thus, the favorable allele can increase in frequency.

In some species, males make a substantial investment in their offspring, and this can affect the pattern of sexual selection. Most dramatic is the nutritional investment in the spermatophore produced by male katydids and other orthopterans. Part of the spermatophore is a sperm-free vessel (spermatophylax) that is eaten by the female after mating. The spermatophore can exceed 30% of the weight of the male, and radiolabel studies show that nutrients from the spermatophylax are transferred to the eggs in the female. This "courtship feeding" is especially important in Mormon crickets, and both female–female competition and male choice have been documented (Gwynne 1984). Studies have shown that the quantity and quality of nutrients in courtship feeding affect the fitness of offspring, but as yet no intraspecific genetic variation has been documented (Gwynne 1988). Nutrient exchange may be a generally important aspect of courtship and mating in insects (Boggs 1981; Rutowski 1983).

Sexual selection as envisaged by Darwin entailed differential mating success, but there is a distinction between differential mating success and assortative mating. In the strict sense, sexual selection results from any process that causes a difference in allele (or genotype) frequency between the pool of potential mates and the pool of successfully mating adults. Once the pool of mating adults is chosen, the genotypes can mate at random, and the situation would still be called sexual selection. On the other hand, with assortative mating, genotypes do not mate at random and mating choices may be made by one or both sexes, but there is not necessarily differential mating success.

Most of the contemporary work on sexual selection, both theoretical and empirical, makes extensive use of quantitative genetic methods. This is because the traits — such as mating behavior, attractiveness, and fighting ability — are complex traits whose transmission usually involves many genes (Lande and Arnold 1985; Kirkpatrick 1982; Kirkpatrick 1985). The discussion of this aspect of sexual selection will be deferred to Chapter 8, where the tools of quantitative genetics are discussed.

The one-locus two-allele model of sexual selection is identical to that of fecundity selection. In particular, imagine a pool of potentially mating adults

that pair off according to weights that determine the relative probability of each genotype pair. These weights, known as **coefficients of sexual selection**, can be expressed as a matrix. With one gene and two alleles, the sexual selection coefficients form a 3 × 3 matrix:

		Male genotype		
		AA	*Aa*	*aa*
Female genotype	*AA*	s_{11}	s_{12}	s_{13}
	Aa	s_{21}	s_{22}	s_{23}
	aa	s_{31}	s_{32}	s_{33}

(4.53)

The most general form of sexual selection allows nine parameters — one for each mating type. (They could be normalized be dividing the matrix by s_{22}, leaving eight "free" parameters.) Since the parameters specify the relative probabilities of different matings, then assuming that each mating produces the same number of offspring, they also represent the relative numbers of offspring from each mating type. Just as special patterns of fecundities were helpful in exploring the dynamics of fecundity selection, the same special patterns can be reinterpreted in terms of sexual selection. One instructive example is when females choose males (and the males differ in desirability as mates). The parameters might be

		Male genotype		
		AA	*Aa*	*aa*
Female genotype	*AA*	$1 - s$	1	$1 - t$
	Aa	$1 - s$	1	$1 - t$
	aa	$1 - s$	1	$1 - t$

(4.54)

The recurrence equations in this example simplify to

$$\bar{s}x' = (x + 1/2y)[(1 - s)x + 1/2y]$$

$$\bar{s}y' = (x + 1/2y)[(1/2y) + (1 - t)z] + [(1 - s)x + 1/2y](1/2y + z) \quad (4.55)$$

$$\bar{s}z' = (1/2y + z)\,[1/2y + (1 - t)z]$$

where $\bar{s}$ is the sum of the right hand sides of the equations. The way the recursions can be factored implies that this form of sexual selection is equivalent in dynamics to viability selection in one sex with no selection in the other sex. You can verify this by noting that the recursions factor into products of gamete frequencies produced by the two sexes. The interior equilibrium for the system of equations is

$$\hat{p} = t/(s + t), \quad \hat{q} = s/(s + t) \quad (4.56)$$

just as with heterozygote superiority, and the equilibrium is stable if $s,t > 0$.

When both sexual selection and fecundity selection are operating, the two matrices of selection coefficients can be combined by forming the element-wise product matrix, and using this product matrix in the above recursions. That is, each element in the product matrix equals the product of the elements at the same position in the original matrices. Fecundity and sexual selection can cancel each other. For example if *aa* males have half the fecundity of the other genotypes, the fecundity matrix is

		Male genotype			
		AA	*Aa*	*aa*	
Female genotype	*AA*	1	1	0.5	
	Aa	1	1	0.5	**(4.57)**
	aa	1	1	0.5	

If these males are also twice as successful in obtaining mates, the fecundity matrix is

		Male genotype			
		AA	*Aa*	*aa*	
Female genotype	*AA*	1	1	2	
	Aa	1	1	2	**(4.58)**
	aa	1	1	2	

Hence, the *aa* males leave as many offspring as other genotypes, because the product matrix is a matrix of 1's, and it would appear as if no selection were operating. The population would remain at Hardy-Weinberg equilibrium for any arbitrary allele frequency.

It should be stressed that there are other ways to formulate models of sexual selection. In one alternative, males and females encounter one another, and males may perform a display of some sort. Females then decide whether or not to mate on the basis of a number of criteria, including the brilliance of the display and the number of previous encounters the female has had. Such models have been considered by Karlin and Raper (1979) and O'Donald (1979, 1980). Box E presents data on differential mating success of the Arctic skua and a swallow, showing the influence of simple genetically determined plumage polymorphisms.

Sexual Selection in Monogamous Birds

E

Polygamous species offer ample opportunity for sexual selection to operate, since the sexually preferred genotype will attract more mates. In monogamous species, however, the significance of sexual selection is somewhat more subtle, since monogamy seems to preclude the attraction of multiple mates. Darwin reasoned that sexual selection can operate in monogamous species if the preferred types mate earlier in the season, and

thereby enjoy increased reproductive success. O'Donald (1987) has been studying the monogamous Arctic skua (*Stercorarius parasiticus*) for many years, and has a good corroboration of Darwin's prediction. There are three color morphs — pale, intermediate, and melanic — that are apparently determined by a single Mendelian gene. The fledging success of skuas declines through the breeding season, and direct observation shows that melanic males tend to breed earlier on average than the others. O'Donald fitted the demographic data to rather involved models that account both for the plumage character and female choosiness, and indirectly explains the maintenance of the polymorphism.

Møller (1988) used a much more direct means of demonstrating female choice in the monogamous swallow *Hirundo rustica*. There is variation in the length of the two conspicuous tail feathers, and by either cutting or gluing extra lengths of feather, he could artificially vary the length of the tail feathers. When the modified birds were released and reproductive success was followed, there was a clear association between tail length and mating success: longer feathered males were at an advantage. Since this swallow is monogamous, the advantage in the reproductive success was seen through the earlier attraction of mates, and the consequent increased likelihood of rearing two broods in a season. This constitutes a good case for selection, but until a clear genetic basis for the trait is established, the evolutionary significance remains unclear. A remarkably similar story was described in *Drosophila melanogaster*: flies caught in the process of copulating had longer wings on average than single flies (Taylor and Kecić 1988).

BALANCE OF OPPOSING FORCES

Most polymorphisms that have any effect on fitness are pleiotropic, necessitating the analysis of models that incorporate more than one selection component simultaneously. Sometimes the pleiotropic effects occur on the same component of selection, but they act in very different ways in the different genotypes. One example is the mechanism whereby β^S/β^A-globin heterozygotes enjoy a heterozygote advantage in West Africa; one homozygote is susceptible to malaria and the other has sickle cell disease. Another striking example of heterozygote advantage mediated by inferiority of both homozygotes through different mechanisms is found among Norway rats in Wales. The poison warfarin is an anticoagulant, and has been used to control rats for a number of years. A single Mendelian gene controls resistance to warfarin, and there is a stable polymorphism for a dominant resistance allele (R). Homozygotes for the wildtype allele (S) are killed by warfarin, and homozygotes for the resistance allele are disfavored because they are highly susceptible to vitamin K deficiency. These two factors combine to give a pronounced advantage to the heterozygotes, to the extent that there are pronounced departures from Hardy-Weinberg proportions in Welsh rat pop-

ulations (Greaves et al. 1977). Polymorphisms that involve such distinct pleiotropic mechanisms may be particularly sensitive to changes in the environment, such as the elimination of malaria or warfarin. Combinations of other selection components are examined in some detail in the following sections.

Mutation–Selection Balance

Recall from Chapter 1 that outcrossing species typically contain a large amount of hidden genetic variability in the form of recessive or nearly recessive deleterious alleles at low frequencies. This situation is to be expected; selection cannot completely eliminate deleterious alleles because of their continual creation through recurrent mutation. To be specific, supposing a is a deleterious allele, consider the model in Table 2, but with mutation allowed. Suppose mutation of A to a occurs at the rate μ per generation. (Since q, the allele frequency of a, will be small, reverse mutation from a to A can be ignored.) The calculation of p' in Table 2 is still valid, except that a proportion μ of A alleles will have mutated to a. Thus

$$p' = p(pw_{11} + qw_{12})(1 - \mu)/\bar{w} \qquad \textbf{(4.59)}$$

Here it is convenient to set $w_{11} = 1$, $w_{12} = 1 - hs$, and $w_{22} = 1 - s$, where $s > 0$. In this context, s is called the **selection coefficient** against aa, and h is the **degree of dominance**; a is completely recessive when $h = 0$, partially dominant when $0 < h < 1$, and dominant when $h = 1$. In the special case $h = 1/2$, the fitness of Aa is exactly intermediate between that of AA and aa and A and a are said to be **additive** in their effects on fitness. In the case of the "dominant" allele for Huntington disease, a severe degenerative disorder of the human neuromuscular system that typically appears after age 35, $w_{11} = 1$, $w_{12} = 0.81$, and $w_{22} = 0$ (data from a Michigan study by Reed and Neel 1959). Thus $s = 1$ and $hs = 1 - 0.81 = 0.19$, so $h = 0.19$. In terms of neuromuscular degeneration, Huntington disease is dominant; in terms of fitness — ability to survive and reproduce — it is only partially dominant due to the late age of onset. Another example is provided by "recessive" lethals in populations of *Drosophila melanogaster*, where the average h is actually in the range 0.025 to 0.050, not 0 (Hiraizumi and Crow 1960; Crow and Temin 1964; Simmons and Crow 1977).

At the equilibrium balance between mutation and selection, $p' = p = \hat{p}$, so

$$\hat{p}(\hat{p}w_{11} + \hat{q}w_{12})(1 - \mu)/\bar{w} = \hat{p} \qquad \textbf{(4.60)}$$

which, after substitution of $w_{11} = 1$, $w_{12} = 1 - hs$, $w_{22} = 1 - s$, and some simplification, leads to

$$\hat{q}hs(1 + \mu - 2\hat{q}) + \hat{q}^2 s = \mu \qquad \textbf{(4.61)}$$

There are two important cases. (1) When a is completely recessive, $h = 0$ and $q^2s = \mu$, so

$$\hat{q} = \sqrt{\mu/s} \tag{4.62}$$

(2) When a is partially dominant, then $\hat{q}$ will often be small enough that $\hat{q}^2s$ can be ignored, and $1 + \mu - 2\hat{q}$ will be close to 1, so $\hat{q}hs \approx \mu$ to a good approximation, or

$$\hat{q} = \mu/hs \tag{4.63}$$

Huntington disease may be used to illustrate application of the above formulas. Recall that for this disorder $s = 1$ and $h = 0.19$. From the expression $\hat{q}hs = \mu$, we could calculate $\hat{q}$ if the value of μ were known, or we could calculate μ if $\hat{q}$ were known. For the Michigan population, $\hat{q} = 0.00005$, and assuming that the population is in equilibrium, we have $\mu = (0.00005)(0.19)(1) = 9.5 \times 10^{-6}$.

As another example, recall from Chapter 1 that the frequency of autosomes containing one or more recessive lethals is about 33% in typical populations of *Drosophila melanogaster* (Ives 1945; Salceda 1977). Thus, the frequency of lethal-free chromosomes is 67%. Suppose that the mutation rate to lethals of any gene is 5×10^{-6} and that a lethal causes a 2.5% reduction in fitness when heterozygous (i.e., $s = 1$ and $h = 0.025$). For such a gene, $\hat{q} = \mu/hs = 5 \times 10^{-6}/0.025 = 0.0002$. For n independent genes, the probability that none will contain a lethal allele is therefore $(1 - \hat{q})^n$; setting $(1 - \hat{q})^n = 0.67$ leads to $n = 2002$, or roughly 2000 genes capable of mutating to lethals. While this estimate of the number of genes in an autosome of *Drosophila* is in good agreement with estimates from other sources, the agreement might be fortuitous, as the model is greatly oversimplified in assuming a constant mutation rate and degree of dominance of each gene.

Figure 25 shows $\hat{q}$ versus μ/s for various values of h. Note that a small amount of dominance, as measured by h, significantly reduces the equilibrium frequency of a; this occurs because when q is small there are so many more heterozygotes than aa homozygotes that even a small amount of selection against heterozygotes has a major effect on the equilibrium frequency of a. Note also in Figure 25 that, for reasonable values of h, μ, and s, $\hat{q}$ is typically less than 0.01. Therefore, while mutation–selection balance satisfactorily accounts for the occurrence of low-frequency detrimentals in populations, it cannot readily account for polymorphisms in which $q > 0.01$.

If there were no mutation, $\hat{q}$ would equal 0 and the average fitness of the population would equal 1. Because mutation occurs, $\hat{q}$ is not 0, and when $h = 0$ the average fitness of the population is

$$1 - q^2s = 1 - (\mu/s)s = 1 - \mu \tag{4.64}$$

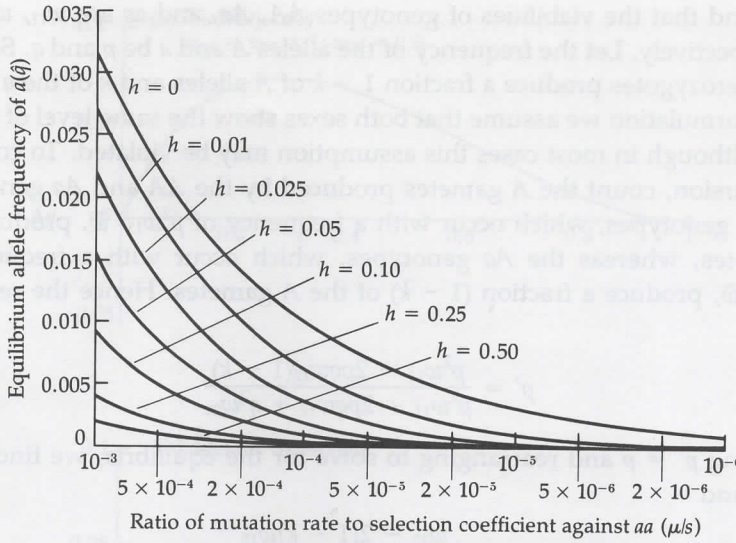


Figure 25. Equilibrium frequency of a detrimental allele with various degrees of dominance (h) against μ/s. The fitnesses of AA, Aa, and aa are assumed to be 1, $1 - hs$, and $1 - s$, respectively; μ is the mutation rate of A to a; and q is the frequency of the a allele.

The reduction in average fitness due to mutation equals $1 - (1 - \mu) = \mu$, which is called the **mutation load**. When a is partially dominant, the mutation load is approximately 2μ. The effect of recurrent mutation in reducing population fitness is therefore independent of how harmful the mutation is, as measured by s. The reason is simply that for a given mutation rate, strongly deleterious mutations have a large s and a low equilibrium frequency, while mildly deleterious mutations have a higher equilibrium frequency. The effect of mutation on population fitness depends only on the mutation rate, so the harmful effect of an increase in mutation rate is the same, irrespective of whether the mutations produced are mildly or severely detrimental. This principle is often called the **Haldane-Muller principle**.

Meiotic Drive and Viability Selection

Incorporation of meiotic drive and viability into one model is motivated by the observation that drive alone cannot maintain a polymorphism, yet there are alleles segregating in natural populations that exhibit meiotic drive. Something must be opposing the meiotic drive, and in the case of the mouse t alleles, the detrimental effects on viability are obvious. Suppose that viability selection occurs by a process of differential survival from zygote to mature

adult, and that the viabilities of genotypes AA, Aa, and aa are w_{11}, w_{12}, and w_{22}, respectively. Let the frequency of the alleles A and a be p and q. Suppose that heterozygotes produce a fraction $1 - k$ of A alleles and k of the a alleles. In this formulation we assume that both sexes show the same level of meiotic drive, although in most cases this assumption may be violated. To construct the recursion, count the A gametes produced by the AA and Aa genotypes. The AA genotypes, which occur with a frequency of $p^2w_{11}/\bar{w}$, produce only A gametes, whereas the Aa genotypes, which occur with a frequency of $2pqw_{12}/\bar{w}$, produce a fraction $(1 - k)$ of the A gametes. Hence the recursion for p is

$$p' = \frac{p^2w_{11} + 2pqw_{12}(1 - k)}{p^2w_{11} + 2pqw_{12} + q^2w_{22}} \tag{4.65}$$

By setting $p' = p$ and rearranging to solve for the equilibria, we find $\hat{p} = 0$, $\hat{p} = 1$, and

$$\hat{p} = \frac{w_{22} - 2(1 - k)w_{12}}{w_{11} - 2w_{12} + w_{22}} \tag{4.66}$$

When $k = 1/2$, segregation is Mendelian and the recursion and equilibria are the same as in the pure viability selection model.

The stability conditions of the fixations are obtained by taking the derivative of p' with respect to p and evaluating at $\hat{p} = 0$ and $\hat{p} = 1$ to obtain the eigenvalues. The eigenvalues are

$$\text{For } \hat{p} = 0\text{: } \lambda = 2(1 - k)w_{12}/w_{22} \tag{4.67a}$$

$$\text{For } \hat{p} = 1\text{: } \lambda = 2kw_{12}/w_{11} \tag{4.67b}$$

The eigenvalues have a simple interpretation in the case of Mendelian segregation, because they indicate that heterozygote advantage guarantees instability of the fixations. If $k \neq 1/2$, then one of the eigenvalues becomes smaller and one larger than the Mendelian situation. The polymorphism can be a protected polymorphism if there is sufficient overdominance, or when directional selection and meiotic drive balance each other. However, not all cases of directional selection and opposing meiotic drive lead to a stable balance (Figure 26).

When directional selection and opposing meiotic drive do balance, the conditions for protection of the polymorphism also guarantee stability of the unique interior equilibrium. The eigenvalue of the interior equilibrium is

$$\lambda = \frac{2(1 - k)w_{11}w_{12} - (2kw_{12} - 2w_{11})w_{22}}{w_{11}w_{22} + 4k(k - 1)w_{12}^2} \tag{4.68}$$

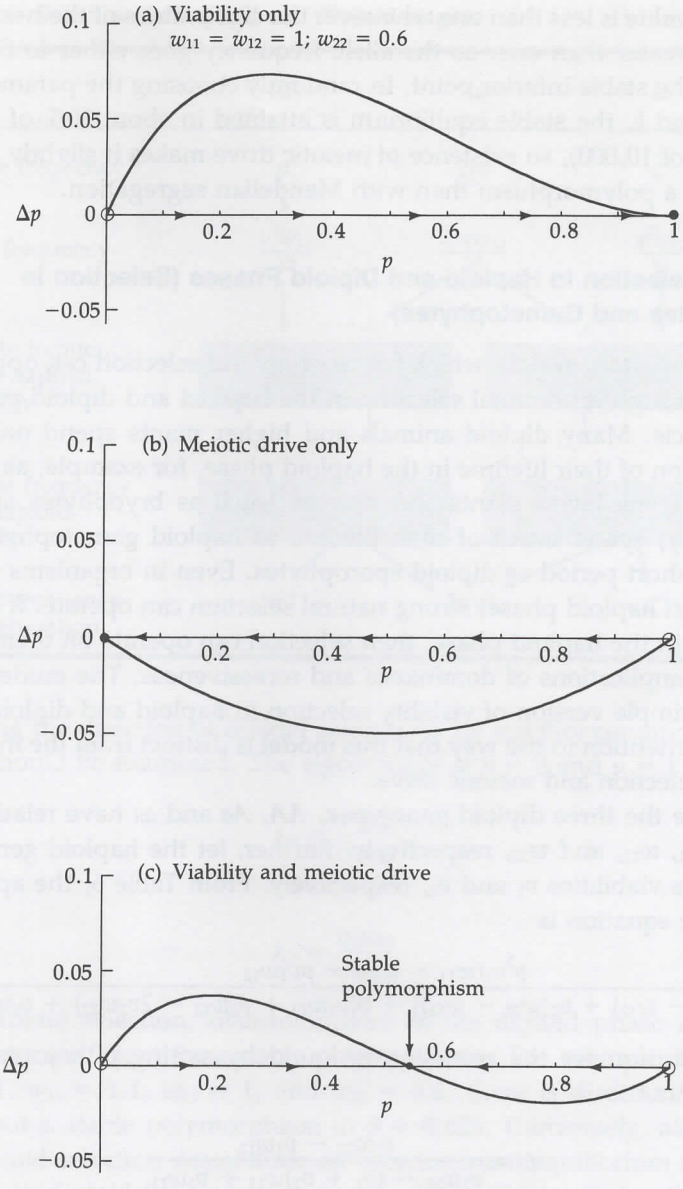


Figure 26. The balance between meiotic drive and viability selection. (a) Δp versus p for viability alone, with fitnesses $w_{11} = w_{12} = 1$, and $w_{22} = 0.6$. In this instance, viability selection would eliminate the *a* allele. (b) Meiotic drive alone, where the heterozygotes *Aa* produce 40% *A* gametes and 60% *a* gametes. With meiotic drive alone, the *A* allele would be lost. (c) Δp versus p when both viability and drive are operating at the same time, using the parameters of panels (a) and (b). In this case, there is a stable polymorphism.

This eigenvalue is less than one whenever the eigenvalues of the two fixations are both greater than one, so the allele frequency goes either to fixation or loss or to the stable interior point. In randomly choosing the parameters w_{11}, w_{12}, w_{22}, and k, the stable equilibrium is attained in about 20% of the cases (1,984 out of 10,000), so existence of meiotic drive makes it slightly *less* likely to observe a polymorphism than with Mendelian segregation.

Viability Selection in Haploid and Diploid Phases (Selection in Sporophytes and Gametophytes)

Another important way in which forces of natural selection can oppose each other is through differential selection in the haploid and diploid portions of the life cycle. Many diploid animals and higher plants spend only a very short portion of their lifetime in the haploid phase, for example, as gametes. However, some lower plants and protists (such as bryophytes and algae, respectively) spend much of their lifetime as haploid gametophytes and a relatively short period as diploid sporophytes. Even in organisms that have only a short haploid phase, strong natural selection can operate. If genes are expressed in the haploid phase, then selection can operate on them directly, without complications of dominance and recessiveness. The model that follows is a simple version of viability selection in haploid and diploid phases. Pay close attention to the way that this model is distinct from the model with viability selection and meiotic drive.

Suppose the three diploid genotypes, *AA*, *Aa* and *aa* have relative viabilities of w_{11}, w_{12}, and w_{22}, respectively. Further, let the haploid genotypes *A* and *a* have viabilities v_1 and v_2, respectively. From Table 6, the appropriate recurrence equation is

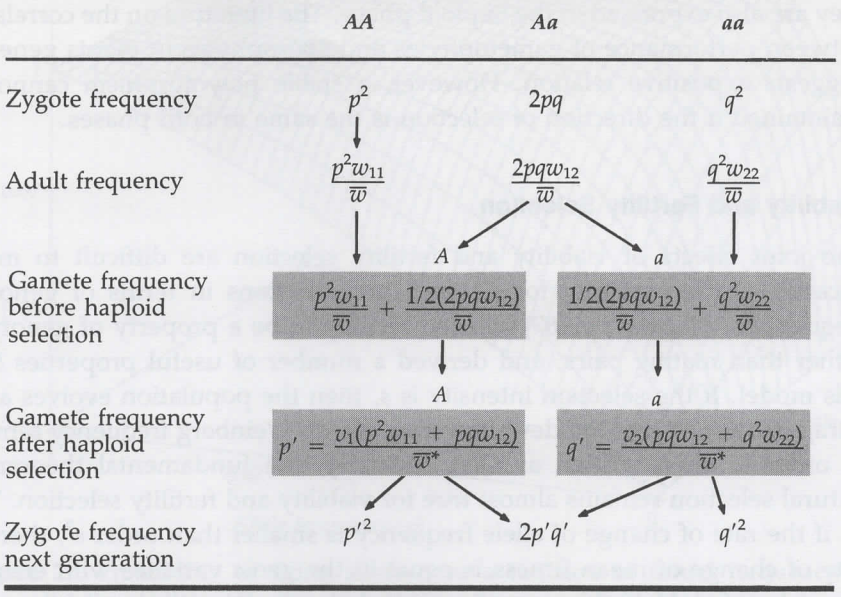
$$p' = \frac{p^2 v_1(w_{11} - w_{12}) + p v_1 w_{12}}{p^2[v_1(w_{11} - w_{12}) + v_2(w_{22} - w_{12})] + p(v_1 w_{12} + v_2 w_{12} - 2v_2 w_{22}) + v_2 w_{22}} \quad \textbf{(4.69)}$$

The equilibrium for the recursion is found by setting $p' = p$. After some algebra,

$$\hat{p} = \frac{v_2 w_{22} - v_1 w_{12}}{v_2 w_{22} - (v_1 + v_2) w_{12} + v_1 w_{11}} \quad \textbf{(4.70)}$$

If $v_1 = v_2 = 1$, the equilibrium is the familiar result from the diploid viability model. Including selection at the haploid stage moves the equilibrium point, but it is still a unique interior equilibrium. When the classical diploid equilibrium exists, the haploid selection may be strong enough to eliminate the equilibrium, in which case the population must evolve to fixation. Thus, haploid selection can destabilize a diploid polymorphism. To determine whether haploid selection can help maintain a polymorphism (for instance,

Table 6. Selection in haploid and diploid stages.



by opposing the direction of diploid selection), the stability conditions of the fixations should be examined. The eigenvalues at $p = 0$ and $p = 1$ are

$$\lambda_0 = \frac{v_1 w_{12}}{v_2 w_{22}} \qquad \textbf{(4.71a)}$$

and

$$\lambda_1 = \frac{v_2 w_{12}}{v_1 w_{11}} \qquad \textbf{(4.71b)}$$

With haploid selection, overdominance in the diploid phase is neither necessary nor sufficient for a stable polymorphism. For example, with $v_1 = 0.8$, $v_2 = 1$, $w_{11} = 1.1$, $w_{12} = 1$, and $w_{22} = 0.6$, there is directional diploid selection but a stable polymorphism at $\hat{p} = 0.625$. Conversely, almost any strong haploid selection destabilizes an overdominant equilibrium when the differences in diploid fitnesses are not very large. To determine the size of the parameter space that admits a stable polymorphism, we analyzed 10,000 random sets of five viabilities for the existence of a stable polymorphism. In 2,082 cases there was a stable polymorphism. Since a diploid viability model would yield about 33% of the cases with a stable polymorphism, the region of the parameter space admitting a stable polymorphism is substantially smaller with diploid and haploid selection than with diploid viability selection alone.

An important point is that haploid selection is more efficient than diploid selection, so traits that are favored in diploids are more efficiently selected if they are also expressed in the haploid phase. The literature on the correlation between performance of gametophytes and sporophytes in plants generally suggests a positive relation. However, a stable polymorphism cannot be maintained if the direction of selection is the same in both phases.

Viability and Fertility Selection

The joint effects of viability and fertility selection are difficult to model because of the necessity for putting the recursions in terms of genotype frequencies. Nagylaki (1987) allowed fertility to be a property of genotypes rather than mating pairs, and derived a number of useful properties from this model. If the selection intensity is s, then the population evolves along a trajectory such that the deviation from Hardy-Weinberg frequency remains of order s, often written as $O(s)$. Similarly, the fundamental theorem of natural selection remains almost true for viability and fertility selection. That is, if the rate of change of allele frequency is smaller than order s^2, then the rate of change of mean fitness is equal to the genic variance with error on the order of s^2. Here the mean fitness is defined as the product of the mean fertility and the mean viability, and mean viability is the geometric mean of the arithmetic mean in the two sexes. Finally, Nagylaki was able to show that with fertility acting as a function of mating pairs, the rate of change of mean fitness approximately equals the genic variance (where mean fitness is defined as the product of the mean viability and the square root of the mean fertility).

Viability Selection in Finite Populations

Despite the heated arguments of the 1970s concerning the relative importance of natural selection and random genetic drift, no population is really infinite in size, so that even with strong selection, sampling effects must be considered. In the case of overdominance, random genetic drift increases the likelihood of fixation. Looking at the problem from the other side, by incorporating overdominant selection in a model of pure drift, the expected time to fixation can actually be decreased by selection. This is because selection can rapidly move the allele frequencies to an equilibrium near fixation, and drift would then require less time to move the rest of way to fixation, providing the equilibrium *is* near fixation, selection is not too strong, and the population size is not too large. With directional selection, drift generally retards the rate of fixation, because drift shifts allele frequencies both up and down.

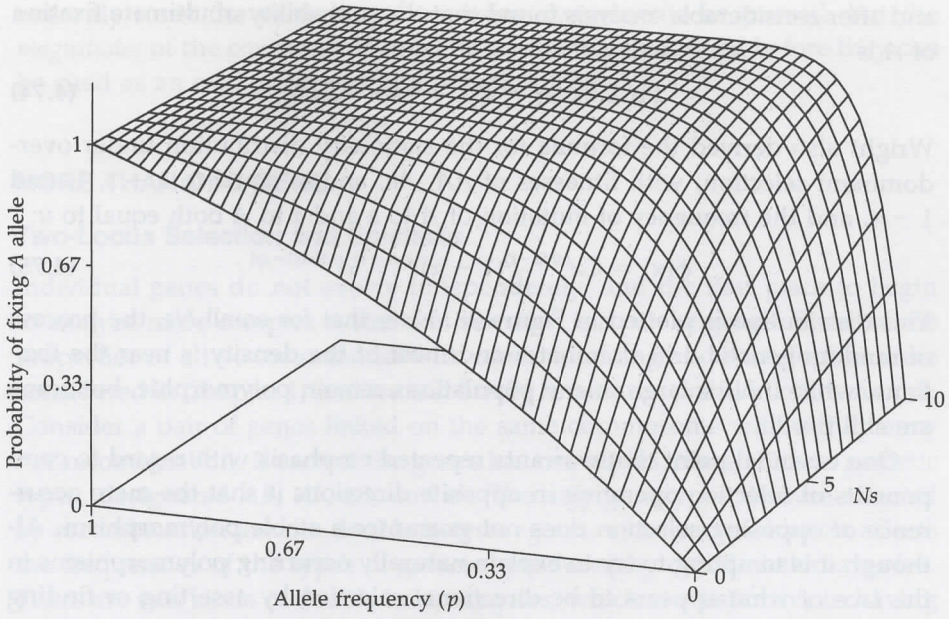


Figure 27. Probability of fixing allele A as a function of initial frequency and Ns under the additive directional selection scheme with fitnesses $w_{11} = 1$, $w_{12} = 1 - s/2$, and $w_{22} = 1 - s$.

With pure drift the fixation probability is p, the initial allele frequency; provided the population is sufficiently large and selection sufficiently strong ($Ns > 1$), the favored allele is more likely to become fixed. However, with $Ns < 0.1$, selection makes little difference, so the fixation probability is close to the initial allele frequency (Figure 27).

In order to model the joint action of random genetic drift and natural selection, Fisher (1958) and Wright (1931) used an extension of the binomial sampling models (the Wright-Fisher model, Chapter 2). Accordingly, let the fitnesses of genotypes AA, Aa, and aa be $1 + s$, $1 + s/2$, and 1, respectively, and let p_i be the frequency of allele A and i be the number of the A alleles. After selection, the allele frequency follows the recursion

$$p_i = \frac{(1 + s)i^2 + (1 + s/2)(i)(2N - i)}{(1 + s)i^2 + 2(1 + s/2)(i)(2N - i) + (2N - i)^2} \qquad \textbf{(4.72)}$$

Fisher and Wright then used the binomial transition probability

$$p_{ij} = \binom{2N}{j} p_i^j (1 - p_i)^{(2N-j)} \qquad \textbf{(4.73)}$$

and after considerable analysis found that the probability of ultimate fixation of A is

$$\Pr(A \text{ fixes}) = s/[1 - \exp(-2Ns)] \tag{4.74}$$

Wright also derived the formula for the stationary distribution under overdominant selection, with fitnesses of AA, Aa, and aa equal to $1 - s$, 1, and $1 - s$, and the frequency of mutation of A to a and a to A both equal to u:

$$\phi(x) = Cx^{(4Nu-1)}(1 - x)^{(4Nu-1)}e^{[4Nsx(1-x)]} \tag{4.75}$$

This distribution is plotted in Figure 28. Note that for small Ns, the process of random genetic drift dominates and most of the density is near the fixations, while if Ns is large, many populations remain polymorphic, but some are still fixed.

One essential point that warrants repeated emphasis with regard to components of selection operating in opposite directions is that the mere occurrence of opposing selection does not guarantee a stable polymorphism. Although it is tempting to try to explain naturally occurring polymorphisms in the face of what appears to be directional selection by asserting or finding evidence of an opposing force, we have just seen specific examples of opposing selection components that nevertheless result in fixation. In real cases,

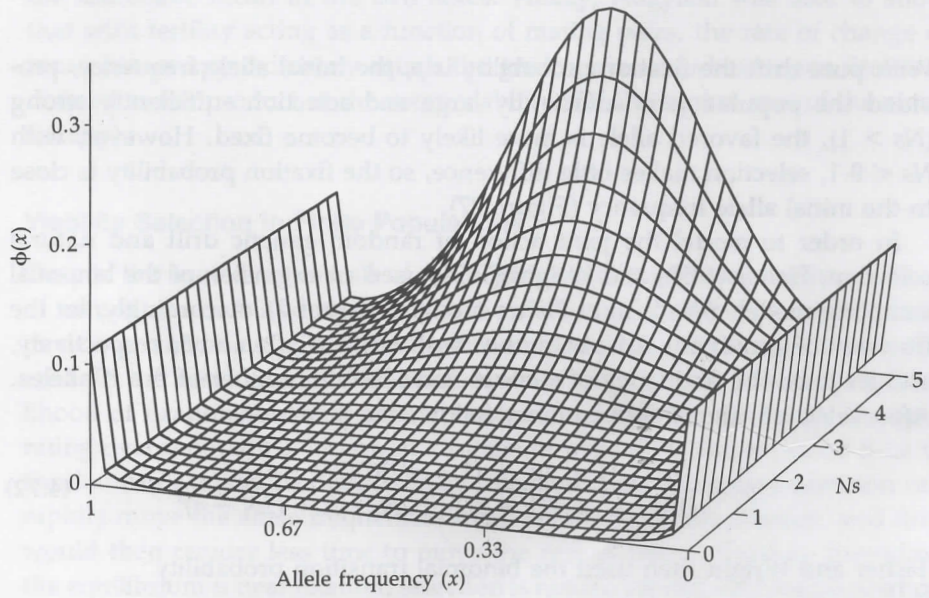


Figure 28. Fraction of populations having allele frequency x [denoted $\phi(x)$] with drift and overdominant viability selection. Fitnesses are $1 - s$, 1, and $1 - s$ for genotypes AA, Aa, and aa, respectively.

not only must the existence of opposing components be proved, but the *magnitudes* of the components must be adequately quantified before they can be cited as an explanation for the polymorphism.

MORE THAN ONE GENE

Two-Locus Selection and Epistasis

Individual genes do not evolve independently, and the first place to begin to analyze more complex systems of selection is a pair of genes. Some of the properties of a two-locus model with random mating and no selection were considered in Chapter 1, and we will briefly review the notation of the model. Consider a pair of genes linked on the same chromosome with a frequency r of recombination. If each of the genes has two alleles, there are four gametic types, designated AB, Ab, aB, and ab. The only genotypes that are affected by recombination are the double heterozygotes, since recombination affects the frequencies of the types of gametes produced by double heterozygotes. There are two kinds of double heterozygotes, which are distinctive in having the alleles in different **linkage phase**. The double heterozygotes are AB/ab, which is called the *cis* or *coupling* phase, and Ab/aB, which is called the *trans* or *repulsion* phase. Table 7 presents the ten possible genotypes and the frequencies of the gametes produced by each.

A slightly more compact notation for two-locus gametes than was used in Chapter 1 gives the four gametes AB, Ab, aB, and ab the respective frequencies x_1, x_2, x_3, and x_4, where

$$\sum_i x_i = 1 \tag{4.76}$$

Let the allele frequencies be

$$\begin{aligned} p_1 &= x_1 + x_2 \\ p_2 &= 1 - p_1 = x_3 + x_4 \\ q_1 &= x_1 + x_3 \\ q_2 &= 1 - q_1 = x_2 + x_4 \end{aligned} \tag{4.77}$$

The viabilities can be specified by the matrix

	AB	Ab	aB	ab
AB	w_{11}	w_{12}	w_{13}	w_{14}
Ab	w_{21}	w_{22}	w_{23}	w_{24}
aB	w_{31}	w_{32}	w_{33}	w_{34}
ab	w_{41}	w_{42}	w_{43}	w_{44}

(4.78)

or, if the two double heterozygotes have the same fitness (i.e., $w_{23} = w_{14}$), then the viability matrix becomes

$$\begin{array}{cccc} & AA & Aa & aa \\ BB & w_{11} & w_{13} & w_{33} \\ Bb & w_{12} & w_{14} & w_{34} \\ bb & w_{22} & w_{24} & w_{44} \end{array} \tag{4.79}$$

Table 7 shows how the recursions are derived from the table of genotypes and gametes they produce. The mean fitness is

$$\bar{w} = \sum_i \sum_j w_{ij} x_i x_j \tag{4.80}$$

The 16 terms in this sum are formed by all possible products of the four gametes taken in pairs, with each multiplied by the fitness of the resulting genotype. The mean fitness is the average of the fitnesses of all genotypes

Table 7. Gamete production table for two-locus viability selection.

	Genotype	Frequency	Fitness	GAMETES PRODUCED			
				AB	*Ab*	*aB*	*ab*
1	*AB/AB*	x_1^2	w_{11}	1	0	0	0
2	*AB/Ab*	$2x_1x_2$	w_{12}	1/2	1/2	0	0
3	*AB/aB*	$2x_1x_3$	w_{13}	1/2	0	1/2	0
4	*AB/ab*	$2x_1x_4$	w_{14}	$(1/2)(1 - r)$	$(1/2)r$	$(1/2)r$	$(1/2)(1 - r)$
5	*Ab/Ab*	x_2^2	w_{22}	0	1	0	0
6	*Ab/aB*	$2x_2x_3$	w_{23}	$(1/2)r$	$(1/2)(1 - r)$	$(1/2)(1 - r)$	$(1/2)r$
7	*Ab/ab*	$2x_2x_4$	w_{24}	0	1/2	0	1/2
8	*aB/aB*	x_3^2	w_{33}	0	0	1	0
9	*aB/ab*	$2x_3x_4$	w_{34}	0	0	1/2	1/2
10	*ab/ab*	x_4^2	w_{44}	0	0	0	1

$$x_1' = \frac{[w_{11}x_1^2 + w_{12}x_1x_2 + w_{13}x_1x_3 + (1 - r)w_{14}x_1x_4 + rw_{23}x_2x_3]}{\bar{w}}$$

$$= \frac{[x_1(x_1w_{11} + x_2w_{12} + x_3w_{13} + x_4w_{14}) - rw_{14}x_1x_4 + rw_{14}x_2x_3]}{\bar{w}}$$

$$= \frac{x_1\bar{w}_1 - rw_{14}D}{\bar{w}}$$

in the population, and the equation comes directly from the assumption that zygotes are formed by random union of gametes. The marginal fitness can be defined as it was for the multiple allele model.

$$\overline{w}_i = \sum_j w_{ij}x_j \tag{4.81}$$

Note that the marginal fitness is not a property of any one gene, but rather includes frequencies of each of the four gametic types.

A useful analogy for two-locus selection is the one-gene, four-allele model, in which the recursion is

$$p_i' = p_i w_i / \overline{w} \tag{4.82}$$

and, from Table 7, we obtain the two-locus recursion

$$x_i' = x_i w_i / \overline{w} \pm r w_{14} D / \overline{w} \tag{4.83}$$

where $D = x_1x_4 - x_2x_3$ is the coefficient of linkage disequilibrium, and the sign is + for $i = 2,3$ and − for $i = 1,4$. If the second term of the two-locus model could be ignored, the two models would be identical. As intuition would suggest, if $r = 0$, the two models are equivalent, because we can label the AB gamete as allele A_1, the Ab gamete as allele A_2, the aB gamete as allele A_3, and the ab gamete as allele A_4. Each "allele" is actually a pair of genes, but in the absence of recombination, they behave as single alleles. Another case in which the two recursions are identical occurs when $D = 0$. With $D = 0$, regardless of the magnitude of recombination, the transition to the next generation is determined by the same equation as in the multiple allele case. This means that a gamete increases or decreases in frequency according to whether its marginal fitness is greater than or less than the mean fitness. After the initial generation, there may be changes in the gamete frequencies to make $D \neq 0$, and then the full recursion must be considered and recombination taken into account.

From a given initial array of gamete frequencies, the four-allele recursion ($r = 0$) predicts a new array for the following generation. Now consider the consequences of a small amount of recombination. Recombination pushes the population slightly away from the $r = 0$ path each generation (Figure 29). This logic would suggest that, by incorporating recombination, many of the principles that were derived for the multiple allele model might still hold approximately. In particular, perhaps mean fitness would be nondecreasing, and the equilibria with complete polymorphisms might be unique (Turner 1970, 1981). However, as we shall see, both of these principles are violated in the two-locus model.

The two-locus selection problem has not been solved for the general case.

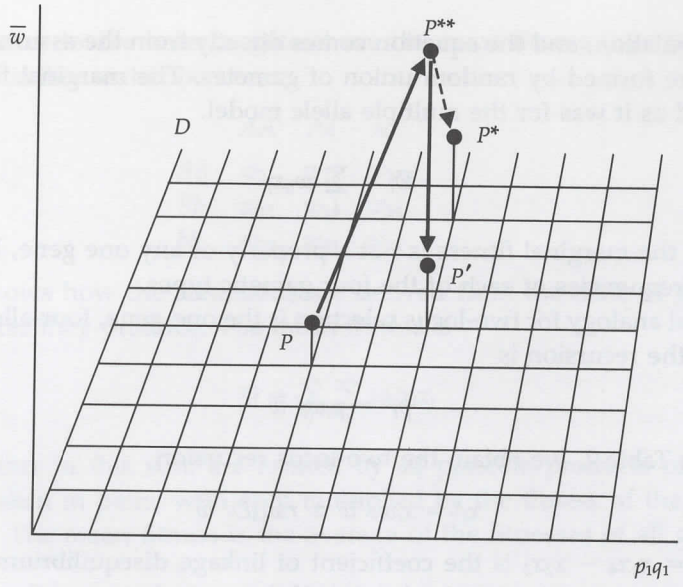


Figure 29. Geometric representation of one generation of selection and mating for two loci. The population starts at point P, is moved by selection to point P^{**}, and by a round of mating with recombination to point P'. If there were no recombination, P^{**} would be moved instead to point P^*, which has the same gene frequencies as point P', because recombination affects D but not the gene frequencies. (From Turner 1981.)

That is, for any arbitrary fitness matrix, there is no simple formula for the equilibria and their stability. However, several special cases with simple patterns of selection have been analyzed. The first is the additive model, which assumes that the fitness effects of the two loci are simply summed to give the fitness of each genotype.

	AA	Aa	aa
BB	$\alpha_{11} + \beta_{11}$	$\alpha_{12} + \beta_{11}$	$\alpha_{22} + \beta_{11}$
Bb	$\alpha_{11} + \beta_{12}$	$\alpha_{12} + \beta_{12}$	$\alpha_{22} + \beta_{12}$
bb	$\alpha_{11} + \beta_{22}$	$\alpha_{12} + \beta_{22}$	$\alpha_{22} + \beta_{22}$

(4.84)

These fitnesses are substituted for the fitnesses w_{ij} into the recursions and solved by setting $x_1' = x_1$, $x_2' = x_2$, $x_3' = x_3$, and $x_4' = x_4$. At equilibrium,

$$\hat{D} = 0$$

$$\text{freq}(A) = \hat{p}_1 = \frac{\alpha_{12} - \alpha_{22}}{2\alpha_{12} - \alpha_{11} - \alpha_{22}} \qquad \textbf{(4.85)}$$

$$\text{freq}(B) = \hat{q}_1 = \frac{\beta_{12} - \beta_{22}}{2\beta_{12} - \beta_{11} - \beta_{22}}$$

Since $\hat{D} = 0$, the frequencies of the gametes at equilibrium are given by

$$\begin{aligned} \hat{x}_1 &= \hat{p}_1\hat{q}_1 \\ \hat{x}_2 &= \hat{p}_1\hat{q}_2 \\ \hat{x}_3 &= \hat{p}_2\hat{q}_1 \\ \hat{x}_4 &= \hat{p}_2\hat{q}_2 \end{aligned} \tag{4.86}$$

Therefore, the additive model gives the same equilibrium as if the two genes were evolving independently. The allele frequencies are determined by the fitness effects of each gene independently, and at equilibrium there is no statistical association between the genes ($\hat{D} = 0$). However, the trajectory of allele frequencies in the two-locus additive model is not necessarily the same as those for single genes. We saw in Chapter 1 that linkage disequilibrium decays at the rate $1 - r$ per generation, and linkage disequilibrium decays over a period of generations with selection in this additive model as well. Indeed, the rate of decay of linkage disequilibrium is *greater* than with no selection, even when there is heterozygote advantage for both genes. The reason for the increased rate of decay is that heterozygote advantage results in an excess of heterozygotes over the Hardy-Weinberg proportions, and the excess of double heterozygotes gives recombination greater opportunity to break down the linkage disequilibrium.

Under the model of additive fitnesses, the mean fitness can be rewritten as

$$\begin{aligned} \bar{w} = \alpha_{11}(x_1 + x_2)^2 + 2\alpha_{12}(x_1 + x_2)(x_3 + x_4) + \alpha_{22}(x_3 + x_4)^2 + \\ \beta_{11}(x_1 + x_3)^2 + 2\beta_{12}(x_1 + x_3)(x_2 + x_4) + \beta_{22}(x_2 + x_4)^2 \end{aligned} \tag{4.87}$$

Written in this manner, the mean fitness is clearly a function of allele frequencies and is independent of linkage disequilibrium because of the identities $p_1 = x_1 + x_2$, $p_2 = x_3 + x_4$, $q_1 = x_1 + x_3$ and $q_2 = x_2 + x_4$. The recursion for the frequency of the A allele can be obtained by adding the recursions for x_1 and x_2:

$$p_1' = (x_1 + x_2)' = (x_1w_1 + x_2w_2)/\bar{w} \tag{4.88}$$

The recursions of allele frequency are also independent of linkage disequilibrium and of the rate of recombination. Since the recursions are independent of r, they must be the same as when $r = 0$, which is the one-locus four allele model. Because allele frequencies change as in the one-locus model, and $\bar{w}$ is a function only of the allele frequencies, it follows that mean fitness is nondecreasing under the additive model. This elegant line of reasoning is due to Ewens (1969, 1979). A contour map of the adaptive topography for a

two-locus additive viability model is given in Figure 30. Note that, once the fundamental theorem has been proven, the global stability of the internal equilibrium is immediately evident.

Another measure of fitness that is useful in understanding the two-locus model is called **induced fitness**. The induced fitness of genotype *AA* is the weighted average fitness of genotypes *AABB*, *AABb* and *AAbb*. Induced fitnesses depend on genotype frequencies, so they change as the system evolves. The definition of the induced fitness of *AA* is

$$w_{AA} = \frac{\text{freq}(AB/AB)w_{11} + \text{freq}(AB/Ab)w_{12} + \text{freq}(Ab/Ab)w_{22}}{\text{freq}(AB/AB) + \text{freq}(AB/Ab) + \text{freq}(Ab/Ab)} \qquad (4.89)$$

where the denominator is the overall frequency of genotype *AA*.

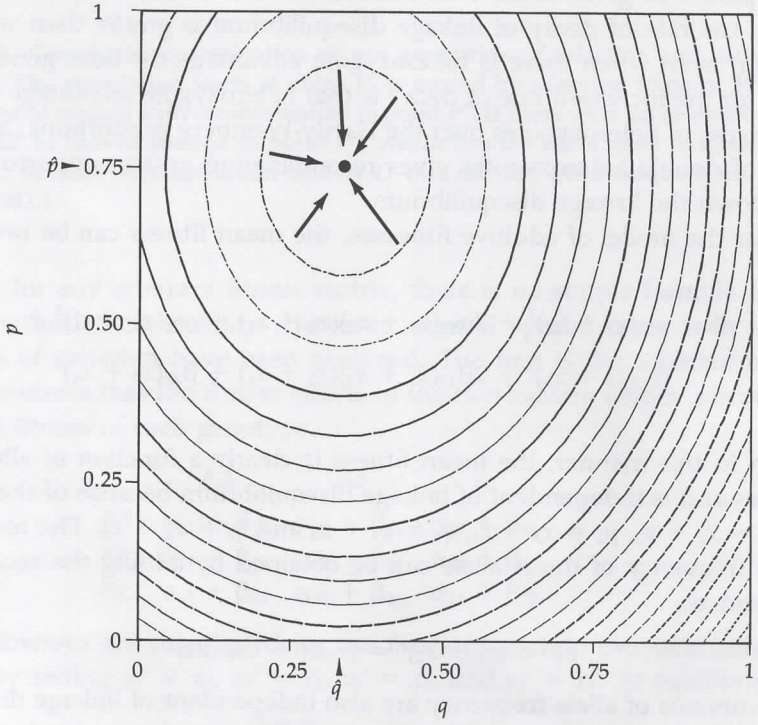


Figure 30. Contour map of adaptive topography for the two-locus additive model. The contours represent the mean fitness of the population for the respective allele frequencies with $D = 0$. The fitness parameters are $\alpha_{11} = 0.9$, $\alpha_{12} = 1$, and $\alpha_{22} = 0.7$ for the *A* locus, and $\beta_{11} = 0.6$, $\beta_{12} = 1$, and $\beta_{22} = 0.8$ for the *B* locus. The dot represents the maximum mean fitness, and in this case it is a globally stable equilibrium with $\hat{p} = 0.75$, $\hat{q} = 0.333$, and $\hat{D} = 0$.

Several other special patterns of two-locus fitnesses have been analyzed in some detail, including a model with multiplicative fitnesses and one with a symmetric pattern of fitnesses (Lewontin and Kojima 1960; Karlin and Feldman 1970). Through the analyses of these special cases, and by examining stability of the boundaries in the general model, some summary conclusions can be made (Karlin 1975; Karlin and Carmelli 1975).

1. The smaller the value of r, the more likely $D \neq 0$ at equilibrium.
2. Nonadditivity of fitnesses (epistasis) is needed for $\hat{D} \neq 0$, and either epistasis must be strong or linkage tight for appreciable linkage disequilibrium (Hastings 1986b).
3. Although $\overline{w}$ usually increases, it does not always (Moran 1964).
4. Overdominance in induced fitness is neither necessary nor sufficient to guarantee a complete polymorphism (Hastings 1982).
5. At least four polymorphisms can be simultaneously stable (Hastings 1985a). See Figure 31 for a view of the adaptive landscape.

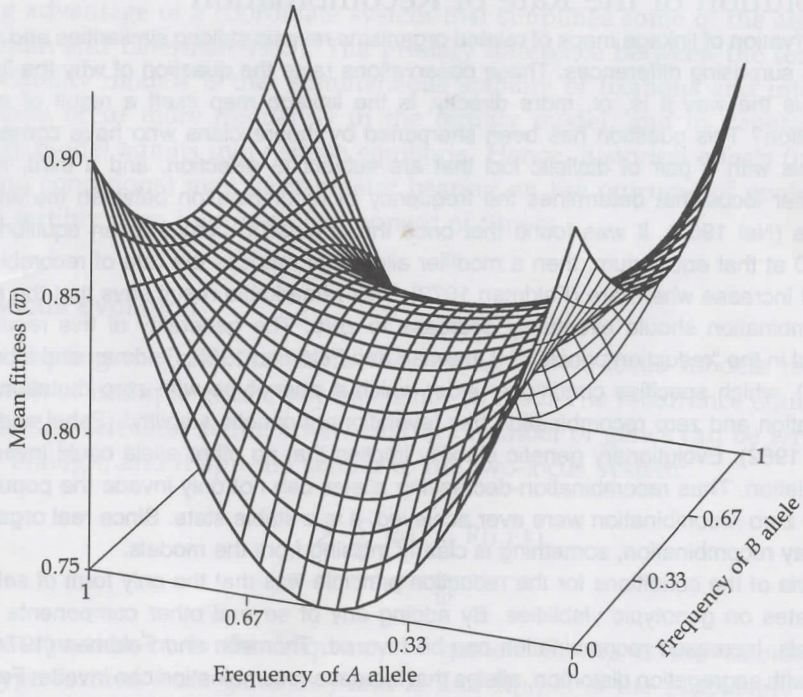


Figure 31. Adaptive topography of two-locus symmetric model that has four stable interior (polymorphic) equilibria. The equilibria are neither maxima nor minima of $\overline{w}$. (Parameters from Hastings 1984.)

Much of the work in two-locus theory that allows deviations from additivity or multiplicativity generally places no limits on the strength of the epistasis. Experimental data, reviewed by Simmons and Crow (1977), Hedrick et al. (1978), and Barker (1979), show that while epistasis in viabilities may exist, it is generally weak. This observation prompted Hastings (1985b) to examine models that allow overdominance but only very weak epistasis. He used an additive model with epistasis parameter ϵ augmenting the homozygote fitnesses, so when $\epsilon = 0$ the model is purely additive. A key result is that when the purely additive model has an interior equilibrium, then for small enough ϵ (i.e., for small enough epistasis), there exists an equilibrium close to that of the purely additive model. If the fitnesses are changed by adding epistasis of order ϵ, then the dependence of equilibrium mean fitness and allele frequencies on the frequency of recombination is of order r^2. This implies that the equilibrium mean fitness and allele frequencies in the additive model are fairly robust to weak epistasis. However, the magnitude of linkage disequilibrium is proportional to $1/r$, with error on the order of ϵ.

F Evolution of the Rate of Recombination

Observation of linkage maps of related organisms reveals striking similarities and sometimes surprising differences. These observations raise the question of why the linkage map is the way it is, or, more directly, is the linkage map itself a result of natural selection? This question has been sharpened by theoreticians who have constructed models with a pair of diallelic loci that are subject to selection, and a third, neutral modifier locus that determines the frequency of recombination between the first two genes (Nei 1967). It was found that once the population comes to an equilibrium, if $D \neq 0$ at that equilibrium, then a modifier allele that reduces the rate of recombination could increase when rare (Feldman 1972). This remarkable result says that the rate of recombination should eventually decrease to zero. The generality of this result was stated in the "reduction principle" (Liberman and Feldman 1986; Feldman and Liberman 1986), which specifies conditions under which a phenotype with zero mutation, zero migration and zero recombination has "evolutionary genetic stability" (Eshel and Feldman 1982). Evolutionary genetic stability implies that no other allele could invade the population. Thus recombination-decreasing alleles can not only invade the population, but if zero recombination were ever achieved, it is a stable state. Since real organisms display recombination, something is clearly missing from the models.

One of the conditions for the reduction principle was that the only form of selection operates on genotypic viabilities. By adding any of several other components to the models, increased recombination can be favored. Thomson and Feldman (1974) saw that with segregation distortion, alleles that increase recombination can invade; Feldman et al. (1980) found that in a mutation-selection balance, recombination-increasing alleles can invade; Felsenstein (1974) saw that in a finite population, recombination rates could increase; and Holsinger and Feldman (1983a) found that in models with a combination

of random mating and selfing, the recombination rate could increase. The generality of the reduction principle has been stated in terms of "perfect" transmission of selected genes having evolutionarily genetic stability (Altenberg and Feldman 1987). Modifier theory promises to be an area of continuing excitement, particularly as mechanisms for modifier function are elucidated. Modification of recombination rate bears directly on a major unsolved problem in evolutionary biology: the evolution of sex (see Bell 1982; Michod and Levin 1987).

Given the complexities of two-locus viability models, fertility models with their added dimensionality might appear completely intractable. A two-locus fertility model would require 10 nonlinear difference equations: one for each genotype frequency. This degree of complexity is unfortunate, because virtually all of our understanding of two-locus evolution is based on a model in which selection occurs only at the viability stage, while there is a large body of experimental evidence that points to fertility as a more important selection component. The key issue is whether two-locus fertility models are qualitatively different from classical viability models. Fortunately, some progress has been made in analyzing the symmetric two-locus fertility model by taking advantage of a coordinate system that simplifies some of the algebra (Feldman and Liberman 1985). The primary difference between the fertility and viability models is that simultaneous stability of fixations and internal equilibria occur more frequently in the fertility model, and in general the fertility model admits more stable equilibria. Hence, historical effects (initial starting conditions) may have greater bearing on the outcome of evolution when fertility is an important component of fitness.

Multilocus Evolution

It is tempting to generalize from the results of two-locus models to the evolution of multiple genes, but this is not so easy. The recurrence equation for viability selection acting on an arbitrary number of genes can be written easily enough, and it appears similar to the two-locus system:

$$x_i' = \frac{\sum_j \sum_k w_{jk} x_j x_k \, \mathrm{R}(i,j,k)}{\overline{w}} \qquad (4.90)$$

In this equation, x_i is the frequency of gamete G_i, w_{jk} is the viability of genotype formed from gametes j and k, and $\mathrm{R}(i,j,k)$ is the probability that genotype G_jG_k produces gamete G_i. Despite the deceptive simplicity of this recursion, no really general result has been obtained from it. On the other hand, a number of properties of special cases have been discovered, and

some comparisons to the two-locus case have been made. The first indication that multilocus models behave differently than two-locus models came from simulations by Lewontin (1964). With overdominance of each locus and sufficiently small values of recombination, he found that the population tended to stabilize with high frequencies of complementary gametes, such as *ABCDE* and *abcde*. One feature of the simulations with five overdominant genes was that genes far apart on the chromosome accumulated substantial linkage disequilibrium, which Lewontin called the cumulative effect. The cumulative effect was explained by arguing that the selection acting on the middle genes generates associations that spread across the chromosome. Another phenomenon observed was the embedding effect, in which adjacent genes exhibit more linkage disequilibrium than in the two-locus model. Clearly, the combined effects of linkage and overdominant selection acting on neighboring genes affect the dynamics substantially. The embedding effect was also explored by Hastings (1984) by using linear programming methods. He found that the strength of selection necessary to maintain given levels of disequilibrium was weaker in a three-locus model than in a two-locus model.

Multilocus simulations were extended by Franklin and Lewontin (1970) by examining models with symmetric multiplicative overdominance (that is, each locus showed overdominance and the fitness of a genotype was determined by the product of the fitnesses of the constituent loci). These simulations revealed a phenomenon the authors called "crystallization," in which complementary gametes achieve high frequency. Crystallization of the genome occurred under conditions of low recombination, strong nonadditive selection, and an initial "nucleus" of allelic association to get it started. As we saw in Chapter 2, genetic drift can cause linkage disequilibrium, and this amount is sufficient to seed the crystallization process. Other multilocus simulations, carried out by Clegg et al. (1976, 1978, 1980), were stimulated by observations that linkage disequilibrium in experimental *Drosophila* populations decays faster than expected by neutrality (Clegg et al. 1976, 1978, 1980). Clegg et al. could reconcile their experimental observations with the simulation results of Franklin and Lewontin (1970) by assuming that the strength of selection was smaller than the simulations, and that there were fewer selected genes per unit of map length. Turelli and Ginzburg (1983) performed multilocus simulations using random viability matrices, and found that fitness generally increases with the number of heterozygous loci at polymorphic equilibria. This is one result that is consistent with simpler models.

Experimental assessment of multilocus interactions using simple segregation tests suggests that high order interactions are uncommon (Clark 1987a). However, this study focused on combinations of entire chromosomes, and epistatic effects of large blocks of genes reflect the average epistatic

effects of the individual genes (Hastings 1986a), so positive and negative interactions may tend to cancel out. Another theoretical approach to multilocus population genetics has been to examine the stability properties of the equilibrium with complete linkage equilibrium. This is known as the **product equilibrium**, because the frequency of any genotype equals the product of the frequencies of the gametes that formed it, and the gamete frequencies are the products of the component alleles. The product equilibrium is stable in the multilocus multiallele additive viability model provided there is some recombination between all genes. Karlin and Liberman (1979, 1982) and Karlin and Avni (1981) have found that, when the product equilibrium is stable for a given linkage map, it is also stable when the frequency of recombination is increased slightly.

The problem of multilocus evolution is deep and complex, and we have only skimmed its surface. Two excellent reviews that will aid your entry into the literature are Hedrick et al. (1978) and Barker (1979).

Multilocus Linkage Disequilibrium

Just as pairwise linkage disequilibrium can reveal selection operating on pairs of genes, so might patterns of statistical association among alleles provide evidence of multilocus selection. We saw in Chapter 1 that available data usually consist of genotype frequencies, while estimates of linkage disequilibrium require estimates of gametic frequencies. Assuming that genotypes are formed by random union of gametes, the two-locus two-allele disequilibrium quantities

$$D = x_1x_4 - x_2x_3 \tag{4.91}$$

and

$$D = x_1 - p_1q_1 \tag{4.92}$$

are identical. With more than two alleles of each gene, there is an analogous linkage disequilibrium statistic for each pair of alleles:

$$D_{ij} = x_{ij} - p_iq_j \tag{4.93}$$

In the case of the HLA (Human Leukocyte Antigen) loci, these pairwise statistics have been very informative. It is also useful to summarize the degree of disequilibrium for each pair of genes, rather than each pair of alleles, and there are several ways to do this (Hill 1975; Weir and Cockerham 1978). The two-allele definition of disequilibrium is readily generalized to three alleles:

$$D_{ijk} = x_{ijk} - p_iq_jr_k \tag{4.94}$$

where x_{ijk} is the observed frequency of gametes containing alleles i, j and k

at the three genes, and the respective allele frequencies are p_i, q_j and r_k. This definition measures the departure of the observed gametic frequency from that expected with independence. However, one problem with these measures is the multiplicity of statistics that become difficult to summarize. Statistically, a reasonable null hypothesis is that the genes are independent, and the method of log-linear analysis can be applied. Log-linear analysis is a powerful technique for testing the statistical dependence of factors in multiway contingency tables, the simplest example being a 2×2 contingency χ^2 test. This approach was first applied to the problem of multilocus genetic data by Smouse (1974). Alternative methods of characterizing multilocus linkage disequilibrium include multivariate analysis of variance (Smouse et al. 1983) and a test based on the variance in the number of heterozygous genes (Brown et al. 1980). The latter test may be particularly applicable to DNA sequence data, since it has no trouble encompassing massive amounts of data, and its power increases with amount of information.

SELECTION COMPONENT ANALYSIS

In this chapter we have been concerned with the theoretical consequences of selection acting on different stages of the life cycle. Results have been obtained by formulating the models in terms of recurrence equations and solving for the equilibria and stability conditions. A quite different problem is to accurately estimate the magnitude of components of selection from samples obtained from natural populations or from experiments performed in the laboratory. In this section we examine one method for analysis of selection components in laboratory experiments, and one method for analysis of selection components using samples from natural populations.

Estimation of Selection Components in *Drosophila*

The use of experimental laboratory populations as a means of determining the relative fitnesses of different genotypes in *Drosophila* has considerable intuitive appeal. In the simplest applications, the procedure amounts to maintaining populations of flies in cages, replenishing the media periodically, and monitoring the genotype frequencies over time. In practice, the method can be much more complex, and inferences drawn from the results have many subtle pitfalls.

Dobzhansky and his co-workers estimated the relative fitnesses of third-chromosome karyotypes in *Drosophila pseudoobscura*. They initiated population cages with known mixtures of the karyotypes and assayed the frequencies of the karyotypes over a number of generations. The selection models

presented previously in this chapter predict the trajectory of gene frequencies for a given array of fitnesses. Here we have the reverse problem: to estimate the fitnesses that give the best fit between the observed trajectories and the theoretical predictions. The fitting method most widely used is the **maximum likelihood** procedure. The idea behind maximum likelihood is to derive a mathematical formula that gives the likelihood of obtaining the observed results (given a particular set of fitness parameters) and then to manipulate the values of the fitness parameters until the likelihood is maximized. In practice, computer programs are used to search for maximum likelihood parameters in efficient and systematic ways. Examples of maximum likelihood methods in fitness estimation are found in Frydenberg (1963), DuMouchel and Anderson (1968), Clark et al. (1981), and Wilson et al. (1982). One problem with the maximum likelihood approach is that many assumptions must be made in deriving the likelihood function, and in the end the estimates are valid only to the extent that the model was correct. For example, one shortcoming of the simple one-locus selection model is that, although it purports to incorporate the net fitness w_{ij} for each genotype, the predicted population trajectory is strictly true only for viability selection. Hence, in the presence of meiotic drive or fecundity selection, maximum likelihood methods that assume viability selection exclusively will yield spurious estimates.

Another problem with estimating fitnesses by fitting population trajectories relates to time of census in the life cycle. Each generation, random mating results in Hardy-Weinberg proportions of zygotes, but selection can result in perturbations from these frequencies. If the census is taken in each generation before all of the selection is complete (that is, additional differential mortality of genotypes or fecundity occurs after census), then the model can be seriously in error (Prout 1965, 1969, 1971a, 1971b; Christiansen et al. 1977). By estimating total fitness in experimental populations and performing separate test crosses to estimate viabilities, fitness differences in viability and fertility have been quantified. Using this procedure, Sved and Ayala (1970) and Sved (1971) determined that fertility components of fitness were often much more important than viability. In an extension of the approach of estimating fitnesses from experimental populations and from separate segregation tests, Seager and Ayala (1982) and Seager et al. (1982) estimated fitness effects of homozygosity for the second and third chromosomes in *Drosophila melanogaster*. This study also revealed important fertility components, and interactions between the chromosomes frequently resulted in double homozygotes having a viability greater than expected. Studies by Anxolabéhère (1980) revealed significant larval selection and frequency-dependent sexual selection among males, and models that accounted for both components gave the best agreement to the changes in allele frequency in 25-generation experimental populations.

The most refined and powerful experimental test of selection components in the laboratory is an approach that fully partitions the components of selection. This approach was followed by Bundgaard and Christiansen (1972), whose protocol is presented in some detail in Box G. Extensions of this design that explicitly estimate the effects of genetic background on fitness components (Clark and Bundgaard 1984), and that estimate two-locus fitness components (Clark and Feldman 1981a) have met with some success. Yet another method of estimating total fitness in *Drosophila* uses competition experiments between a strain of interest and a laboratory strain bearing compound autosomes. Compound autosomes are derived from a pair of metacentric chromosomes, with both left arms attached to a common centromere or both right arms attached to a different centromere. The offspring of the cross between the tested strain and the compound autosome strain are inviable because of chromosome imbalance. The only offspring that survive result from matings in which both parents are either of the tested strain or of the compound autosome strain. Although this approach does not eliminate all the problems besetting laboratory selection experiments, it is a quick and convenient method, with no need for complex scoring methods. The method was used by Jungen and Hartl (1979), and a comparison of fitness estimates obtained by this and other methods showed significant disparities (Haymer and Hartl 1983). Attempts at direct estimation of fitness components have been less common the last few years, largely because of the increased interest in molecular phenomena. Nevertheless, it remains an important problem to relate observed genetic variation to differences in fitness. A review of the subject of fitness estimation using *Drosophila* can be found in Hedrick and Murray (1983).

G Estimation of Selection Components in Experimental Populations of *Drosophila melanogaster*

Using a set of easily scored genetic markers that segregated like one diallelic locus, Bundgaard and Christiansen (1972) devised a series of experiments that allowed them to determine the magnitudes of several selection components. Two hundred virgin males and females of the three genotypes A_1A_1, A_1A_2, and A_2A_2 were placed, in various frequencies, into a flask that served as a mating arena. Most of the females mated within two hours and were transferred into individual vials, where they laid eggs. When the offspring emerged it was possible, knowing the maternal genotype and the array of offspring genotypes, to infer the paternal genotype. If the frequency of alleles among paternal gametes differed from the original paternal gene frequency, this provided evidence for sexual selection in males. Similarly, if the frequency of female genotypes that failed to produce progeny differed from the total female frequency, this implied

sexual selection operating among the females. Deviations from Mendelian ratios among offspring implied either meiotic drive or viability selection. Genotype-pair dependent progeny counts suggested that fecundity selection was operating.

By formalizing the statistical scoring of all of these variables, Bundgaard and Christiansen obtained estimates of Δp for each selection component in a one-generation interval. In the case of the markers *sparkling*, *cubitus interuptus*, and *shaven-naked*, they found that the sexual selection component was by far the greatest, with the multiply heterozygous males having sufficient mating advantage to result in a stable polymorphism.

Estimation of Fitness Effects Using Selective Media

One of the problems that has plagued the investigation of fitness components is that there is no simple way to ascribe the observed effects to any particular target gene. For example, estimates are often sought for the relative fitnesses of alternative genotypes of a single gene, but it is very difficult to construct strains that differ in only one gene. The large selection coefficients that are often ascribed to allozyme genes have raised suspicions about the validity of measuring single-locus fitness in the usual way (Mukai and Yamazaki 1980). One study was designed to quantify the effects of genetic background (Clark and Bundgaard 1984), and the clear overdominance pattern seen in the initial stocks with linkage disequilibrium disappeared as the linkage disequilibrium was allowed to decay. Several methods have been used to get around this problem of genetic control, including a "randomization" of the genetic background by intercrossing. With randomized genetic backgrounds, there is extensive genetic variation outside of the target gene, but under ideal conditions there is no linkage disequilibrium between the target and nearby genes, so that selective effects can be ascribed to the target gene.

Another way to ascribe selective effects to a target gene is with selective media of the sort long in use for the recovery of mutant microorganisms. Application of selective media in population genetics was novel in attempting to address the question of selective neutrality of naturally occurring allozymes. One of the first uses of this technique was in the study of the amylase polymorphism in *Drosophila melanogaster* (DeJong and Scharloo 1976). Amylase hydrolyzes starch into the sugar maltose, and it was thought that flies with high-activity alleles for amylase might be at a selective advantage in medium with large amounts of starch, while the difference might disappear in standard medium rich in simple sugars. A positive result was obtained with *D. melanogaster* (with corroboration by Hickey 1979 and by Clark and Doane 1984) and with *D. pseudoobscura* by Powell and Andjelkovic (1983) and Powell and Amato (1984).

A somewhat less direct association between a selective medium and a target locus was used by Bijlsma and van Delden (1977), who tested the relative fitnesses of genotypes of glucose-6-phosphate dehydrogenase (G6PD) and 6-phosphogluconate dehydrogenase (6PGD) in *Drosophila melanogaster*. These enzymes are the first two enzymes in the pentose phosphate shunt (so called because glucose units are "shunted" into this pathway away from glycolysis). The pathway serves primarily to reduce $NADP^+$ to NADPH; these reducing equivalents are then used in a variety of biosynthetic pathways, including lipogenesis. By supplying a fatty acid, the need for reducing equivalents is lessened, and it was reasoned that high-activity alleles for enzymes of the pentose phosphate shunt would lose any selective advantage they might have. Bijlsma and van Delden (1977) obtained results that were reasonably consistent with the prediction using sodium octanoate as the fatty acid. Later, Eanes et al. (1985a) constructed a series of strains using tightly flanked markers so that different alleles for G6PD could be compared in different media and in different genetic backgrounds. Although their sample size was twice that of the earlier study, they failed to find an effect of fatty acids. Repeatability is a cornerstone of modern science, and the failure to repeat the previous experiments forces a careful reanalysis of the problem. Several explanations can be entertained for the difference between the two results; the most likely one is that the genetic backgrounds were different, but the flat contradiction between the results raises questions about the reliability of the approach.

In terms of control of environmental variables, isogenicity of competing strains, and reproducibility of results, perhaps the cleanest use of selective media to ascribe fitness differences to particular alleles is that of Dykhuizen et al. (1984a). In this experiment, the relative growth rates of strains of *Escherichia coli* that differed in the *gnd* gene (coding for the enzyme G6PD) were estimated in competition against a standard strain for a resource that limited growth. Details of some of the experiments were discussed in Box C.

Estimating Components of Selection Using Samples from Natural Populations

The formal statistical analysis of selection components using mother–offspring samples from a natural population began with a key paper by Christiansen and Frydenberg (1973). The scheme for estimating components of selection was devised for *Zoarces viviparus*, a live-offspring–bearing fish that is found in the Kattegat and Aarhus Bay in Denmark. Samples of adult fish were caught and electrophoretic analysis performed to determine the genotypes of adult males, gravid and nongravid adult females, and a single offspring from each female. The basic idea behind the estimation procedure

was that, under the hypothesis of no selection, predictions could be made about the genotype frequencies in these samples, and selection could be inferred from departures from the expectations. For example, with no selection, gravid and nongravid females should have the same genotype frequencies. Therefore, if the frequency of heterozygotes among gravid females was greater than that among nongravid females, this would imply that heterozygous females were more likely to become gravid, and fertility selection favoring heterozygous females would be suggested. The following describes the formal statistical procedure.

The data from the sample can be tabulated into a form like that in Table 8, showing the counts of all mother–offspring combinations and of male genotypes. A hierarchical series of hypotheses are tested sequentially using a χ^2 goodness-of-fit test for successively simpler models (Table 9).

The first hypothesis tested is whether one-half of the progeny of heterozygous females are heterozygous. Under normal Mendelian segregation of two alleles, this would be expected regardless of the genotype of the father. If this hypothesis is rejected, it implies that meiotic drive occurs in females. The second hypothesis is that male and female gametes unite at random. If

Table 8a. Format of the frequency data from a population for the mother–offspring method of selection component analysis.

	Mother–Offspring Combinations						
	A_1A_1	A_1A_2	A_2A_2	Σ	Sterile females	Adult males	Genotype
A_1A_1	C_{11}	C_{12}	—	F_1	S_1	$A_{\delta 1}$	A_1A_1
A_1A_2	C_{21}	C_{22}	C_{23}	F_2	S_2	$A_{\delta 2}$	A_1A_2
A_2A_2	—	C_{32}	C_{33}	F_3	S_3	$A_{\delta 3}$	A_2A_2

Table 8b. Expected genotype frequencies under random mating and in the absence of selection.

	Mother–Offspring Combinations				
	A_1A_1	A_1A_2	A_2A_2	Σ	Sterile females and adult males
A_1A_1	π_1^3	$\pi_1^2\pi_2$		π_1^2	π_1^2
A_1A_2	$\pi_1^2\pi_2$	$\pi_1\pi_2$	$\pi_1\pi_2^2$	$2\pi_2\pi_2$	$2\pi_1\pi_2$
A_2A_2		$\pi_1\pi_2^2$	π_2^3	π_2^2	π_2^2

(From Christiansen and Frydenberg 1973.)

Table 9. Hierarchical series of hypotheses tested for random mating and selection on the observations.

Hypothesis	Description of component tested	Degrees of freedom	χ^2	P
H_1	Gametic selection in females	1	0.34	>0.50
$H_{02} \mid H_1$	Random mating of breeding fish	2	1.37	>0.50
$H_{03} \mid H_{02}$	Differential male mating success	1	0.98	>0.30
H_4	Differential female mating success	2	0.37	>0.50
H_{04}	Random mating, gametic and sexual selection	6	3.06	>0.80
$H_{05} \mid H_{04}$	Zygotic selection equal in sexes	2	0.22	>0.80
$H_{06} \mid H_{05}$	Zygotic selection	1	0.09	>0.70
$H_{06} \mid H_{04}$	Total zygotic selection	3	0.31	>0.95
H_0	Random mating, gametic, sexual and zygotic selection	9	3.37	>0.90

χ^2 tests are from the data in Table 10.
(From Christiansen and Frydenberg 1973.)

this hypothesis is rejected, it implies nonrandom mating, or perhaps an odd sort of male gametic selection that differs depending on female genotype. The third hypothesis is that the frequency of transmitted male gametes is equal to the frequency of the respective alleles in adult males. If this hypothesis is rejected, it can be due either to male sexual selection (because some males will be involved in more matings than others) or to gametic selection among male gametes. This type of sampling cannot separate these two possibilities, and they are said to be confounded.

The fourth hypothesis is that the genotype frequencies in gravid females are the same as the genotype frequencies in nongravid females. Rejection of this hypothesis indicates female sexual selection. The fifth hypothesis is that zygotic selection, if it occurs, is the same in both sexes. This is tested by a χ^2 test of whether the adult genotype frequencies in males and females are the same. The final hypothesis is that there is no zygotic selection, which is tested by a χ^2 goodness-of-fit test to the Hardy-Weinberg proportions. As discussed in Chapter 1, a goodness-of-fit test to Hardy-Weinberg proportions is not by itself a very powerful test because strong selection of certain types

preserves Hardy-Weinberg proportions (Lewontin and Cockerham 1959). However, testing goodness-of-fit to Hardy-Weinberg proportions *after all other hypotheses* have not been rejected is much more meaningful. Fecundity selection is not directly tested from mother–offspring samples, but separate data are collected on the number of offspring from each female genotype, and standard statistical methods are used to test the null hypothesis that they are the same. An example of the application of selection component analysis, following Christiansen and Frydenberg, is given in Table 10. This example indicates that the polymorphism of esterase alleles in *Zoarces* cannot be explained by the components observed, since the only significant component was an underdominance in viability. In other cases, such as the *Sex ratio* chromosome in *Drosophila pseudoobscura*, selection components analysis has been useful in identifying the forces that act upon a polymorphism (Curtsinger and Feldman 1980; discussed further in Chapter 9). A number of extensions of this approach have also been developed, including the use of multiple successive cohorts (Christiansen et al. 1973, 1977), the joint analysis of more than one gene (Østergaard and Christiansen 1981), and the analysis of multiple offspring per female (Siegismund and Christiansen 1985).

Quantifying the effects of selection in humans presents some unique problems. Other than the frequently cited case of β^S globin, the bulk of natural selection in humans occurs *in utero* in that most genetic disorders are eliminated by fetal loss and stillbirth. One means of inferring the magnitude

Table 10. Observed genotype distribution in the esterase polymorphism of the eelpout.[a]

	Mother–Offspring Combinations					
	A_1A_1	A_1A_2	A_2A_2	Σ	Sterile females	Males
A_1A_1	41	70	—	111	8	54
	(39.0)	(67.0)	—	(106.0)	(9.3)	(58.4)
A_1A_2	65	173	119	357	32	200
	(67.0)	(181.9)	(114.9)	(363.8)	(32.1)	(200.5)
A_2A_2	—	127	187	314	29	177
	—	(114.9)	(197.3)	(312.2)	(27.6)	(172.1)
Σ	106	370	306	782	69	431
	(106.0)	(363.8)	(312.2)	—	—	—

[a] Numbers in parentheses are those expected under random mating in the absence of selection. (From Christiansen and Frydenberg 1973.)

of fetal loss makes use of genetic markers linked to the putatively selected gene. Diabetes is a complex disease with many forms, but the form known as insulin-dependent diabetes mellitus (IDDM) seems to fit a single major gene model rather well, with the gene either linked to or in the HLA region. There is a strong statistical association between HLA alleles *DR3* and *DR4* with IDDM (95% of cases of IDDM have *DR3* or *DR4*). Fathers heterozygous for the *DR4* allele transmit *DR4* to 72.1% of their offspring, while female heterozygotes show normal transmission. *DR3* is transmitted to offspring at an elevated rate (65–70%) by both sexes. Although meiotic drive has not been firmly ruled out as a mechanism for the distorted ratios, it is more likely due to an excess *in utero* mortality of fetuses carrying neither the *DR4* nor the *DR3* allele. This constitutes a case where the genotypes with the highest prenatal viability later suffer from a serious disease, and may account for the relatively high frequency of IDDM. Other associations between HLA alleles and spontaneous abortions are well documented (Cavender et al. 1984).

MOLECULAR VARIATION AND NATURAL SELECTION

Methods for determining the constraints that natural selection imposes on variation in DNA and proteins are discussed in Chapter 7, but it is useful at this point to discuss the distinction between inferences made from molecular data and those of selection component analysis.

One approach to the analysis of molecular variation is to construct hypotheses from neutral mutation theory and test of the hypotheses statistically. Models of strict neutrality assume that every mutation is completely neutral and that the dynamics of allele frequency change are solely due to mutation, random sampling, and perhaps migration. These models make predictions about the relative frequencies of alleles of multiallelic loci. In particular, the neutral-allele sampling theory developed by Ewens (1972) and Watterson (1978) gives the expected allele frequency distribution for any predetermined sample size and number of observed alleles. Statistical tests have also been devised to test goodness-of-fit between the observed and expected allele frequency distributions. Models of strict neutrality also yield predictions about the structure of variation in geographically subdivided populations, and a variety of tests of these predictions are presented in Chapter 7.

Perhaps not surprisingly, models of strict neutrality are often rejected (Keith 1983; Keith et al. 1985; Hedrick and Thomson 1983; Whittam et al. 1986). Rejection does not necessarily imply natural selection, because other explanations for rejection include failure of the population to be at equilibrium with respect to mutation and random genetic drift, hierarchical structure in the population, a growing population size, and other departures from the assumptions of the model. Because goodness-of-fit tests cannot distinguish

among these possibilities, they must not be construed as tests of natural selection.

Another type of data that is becoming increasingly available to the population geneticist allows inference of polymorphism from DNA sequences. DNA sequences are particularly useful when the function of the gene is well characterized and the protein-coding regions (exons) are known. The relative numbers of silent site polymorphisms and amino acid replacement polymorphisms can then be determined and compared. When sequences are compared across diverse taxa, molecular evolutionists often infer that portions of sequences that are conserved are under selective "constraints." This inference derives from the expectation that neutral mutations will accumulate at a more or less constant rate and in a uniform distribution along a sequence. Strong departures from this expectation suggest that, even if mutations occurred uniformly along the sequence (they probably don't; see Chapter 3), some nucleotide changes are likely to render the gene nonfunctional. Such changes are rapidly lost from the population, giving a slower rate of nucleotide substitution at those sites. When DNA sequence variation is known within an entire species, the detailed molecular structure of polymorphisms can be determined, and it is possible to infer selective constraints by a similar argument. For example, Kreitman (1983) sequenced 11 alleles of the gene for alcohol dehydrogenase (*Adh*) in *Drosophila melanogaster*. The coding region contained 765 bp, of which 192 were silent sites (nucleotide positions at which two or more different nucleotides code for the same amino acid). Kreitman (1983) observed 13 silent polymorphic sites within the coding region of the *Adh* gene and only one site that resulted in an amino acid polymorphism (the classic *fast/slow* electrophoretic polymorphism). Selective constraints were inferred to limit the frequency of amino acid polymorphisms simply because amino acid polymorphisms were present at a much lower frequency than would be expected based on the level of polymorphism at silent sites.

An excess of silent substitutions may be good evidence for selective constraints and explain why more polymorphisms do not occur, but it fails to provide a test for the type of selection acting on polymorphisms that are maintained. For example, the conclusion that there are selective constraints on *Drosophila* alcohol dehydrogenases does not say why the *fast/slow* polymorphism exists. Sawyer et al. (1987b) have devised a statistical test of neutrality of amino acid polymorphisms by comparing the allele frequency distribution with that of the silent site polymorphisms in the same gene. When applied to the gene for 6PGD in *E. coli*, the test indicated that no more than half of the amino acid polymorphisms could be selectively neutral. However, the estimated selection coefficients against amino acid replacements was very small, averaging 1.6×10^{-7}.

Along the same lines, Hudson et al. (1987) describe a test that is based on samples of two or more genes from two or more species. The idea is that the degree of divergence between species should be mirrored in the degree of polymorphism within species. They compared the *Adh* coding and flanking regions in *D. melanogaster* and *D. sechellia*. The degree of interspecific divergence was roughly the same for the two regions (the fraction of sites that differed was 0.052 in the coding region and 0.056 in the flanking region), while the coding region of *Adh* was more polymorphic than the flanking region (the fractions of polymorphic sites were 0.101 and 0.022, respectively). The comparison suggests that balancing selection has retained greater than expected levels of polymorphism in the coding region, and formal statistical tests support this conclusion.

Much remains to be learned about how population genetic phenomena can be inferred from sequence data. For example, the arguments above consider sites as independent. Since the sites occur in sequence, the joint information in contiguous sites yields additional information. Joint information can be used to test predictions about linkage disequilibrium, but classical theory of linkage disequilibrium is not directly applicable, because at very low levels of recombination forces such as recurrent mutation and gene conversion play a significant role. Information from DNA sequences can also be used to infer the phylogeny of alleles, and this is an area where interesting tests of neutrality are being devised.

Despite all the excitement about the rigor of tests that can be made using DNA sequence data, it is important to emphasize what cannot be learned from sequences that can be learned from direct tests of selection components. Classical tests of selection components analyze the biology of the organism and partition the life cycle according to mechanistically distinct stages. This can reveal the balance of opposing forces in different life stages. On the other hand, selection component data are always subject to the limitation that enormous sample sizes are needed to obtain adequate statistical power. Molecular data obviate the need for astronomical sample sizes, but at a cost. The cost is the inability to specify the mechanisms of selection. For example, a given set of molecular data may imply that selection was operating in the past, but it does not specify whether the selection was due to viability, fertility, meiotic drive, sexual selection, or any other process.

ECOLOGICAL COMPLICATIONS

The models that have been presented in this chapter can become mathematically complex, but considering the diversity of factors that can affect the

selective differences among genotypes, the theoretical models are simplistic. The models are nevertheless useful in providing a framework upon which to build testable hypotheses, and in providing a starting point for incorporating ecological factors into the models.

One obvious simplification in the models presented in this chapter is the implicit assumption that fitness is an invariant property of genotypes. In the real world, changes in the environment may affect different genotypes in different ways, and may even augment the reproductive fitness of some genotypes while diminishing that of others. Another complication is that the range of a species may be broken into habitats that differ in environmental parameters resulting in different fitnesses for the different genotypes in different places. Relevant models of population subdivision and migration are discussed in Chapter 6. In Chapter 9, temporal and spatial variation in selection coefficients are considered.

With the exception of the model of frequency-dependent selection, the models in this chapter have assumed that fitness is independent of biotic factors such as inter- or intraspecific competition. Models that incorporate these complications are discussed in Chapter 9. Another important aspect that has been ignored is population density — fitness of genotypes might depend on population density. The absolute fitness of genotypes must depend on density to keep population size in check, which may or may not affect relative fitnesses. Density-dependent selection has been modeled by ascribing different density-dependent fitness functions to single-locus genotypes; some results are summarized in Chapter 9. Real populations also have an age structure, and models of natural selection with age-dependent fitness functions yield insight into the evolution of senescence.

Multifactorial evolutionary models consider the phenotype first and attempt to build a statistical description of the changes in phenotype based on underlying assumptions of genetic transmission. Quantitative genetics has had a recent revival among evolutionary biologists, largely because complex and biologically relevant phenotypes can be analyzed. Quantitative genetics allows evolutionary biologists to study traits that have a complex genetic determination, and to relate them to evolutionary theories. Examples of such traits are cold tolerance, pigmentation patterns, horn sizes, and various aspects of dentition. Quantitative genetics can readily incorporate the added complexities of environmental effects by admitting terms of genotype–environment interaction. However, one limitation of this approach is that it is fundamentally a statistical description that cannot identify properties of individuals genes. The strengths and limitations of quantitative genetic models are explored further in Chapter 8.

PROBLEMS

1 In a multiple allele system in which each heterozygote is superior to the homozygotes for the alleles it contains, why are only a subset of alleles usually maintained by selection?

2 If alleles A_1, A_2, A_3, and A_4 are equally frequent and additive in their effects, and the homozygote fitnesses are as follows, what are the heterozygote fitness and $\overline{w}$?

A_1A_1	0.8
A_2A_2	0.6
A_3A_3	0.4
A_4A_4	0.2

3 If the fitnesses of AA, Aa, and aa are 1.0, 0.9, and 0.6, and $p_0 = 0.7$, calculate p_1, p_2, and p_3 for three generations of selection.

4 Calculate the equilibrium allele frequency with overdominance when the fitnesses of AA, Aa and aa are:

a. 0.300, 1, 0.700

b. 0.930, 1, 0.970

c. 0.993, 1, 0.997

5 Calculate $\overline{w}$ for $w_{11} = 0.9$, $w_{12} = 1$, $w_{22} = 0.6$ and $p = 0.8$, assuming random mating. Does any other p give a larger $\overline{w}$? Why or why not?

6 If a rare allele that is lethal when homozygous decreases in frequency by one percent each generation (i.e., $q' = 0.99q$), then what is the selection coefficient against heterozygotes? (Hint: Assume that qh is small compared to 1.)

7 The mutation rate to a dominant gene for neurofibromatosis is approximately 9×10^{-5} and the reproductive fitness of affected individuals is estimated as 1/2. What is the expected equilibrium frequency of affected individuals at birth?

8 What is the equilibrium frequency of a recessive gene arising with a mutation rate of 4×10^{-6} and a reproductive fitness in homozygotes of 0.8? What would it be if the gene were partially dominant with $h = 0.05$?

9 What is the equilibrium frequency of a recessive gene arising with a mutation rate of 10^{-6} with a fitness of 0.4 in homozygotes? How much would this be reduced if the homozygotes did not reproduce at all?

10 For a rare allele maintained at an equilibrium frequency of $\hat{q} = \mu/h$, where h is the selection coefficient against heterozygotes, show that the proportion of heterozygous zygotes resulting from new mutations is approximately equal to h.

11 If selection is not too intense, an additive gene giving fitnesses $1 + s$, $1 + s/2$, and 1 in AA, Aa, and aa will increase in frequency approximately according to $\ln(p_t/q_t) = \ln(p_0/q_0) + (s/2)t$. Calculate the approximate number of generations required to evolve significant insecticide resistance in an insect population when $s = 1/2$ and $p_0 = 10^{-5}$. Significant resistance in the population may be taken as $p_t = 10^{-1}$. Show that, when p_t, $p_0 << 1$, $t \approx (2/s)\ln(p_t/p_0)$.

12 Show that a random mating diploid population with fitnesses 1, $1 - s$, and $(1 - s)^2$ for AA, Aa, and aa gives the same change in the allele frequency p of A as a haploid population with fitnesses 1 and $1 - s$ of A and a.

13 Suppose that in the ith generation of a haploid population the fitnesses of A and a are s_i:1. Letting p_i and q_i be the relative frequencies of A and a, respectively, in generation i, show that

$$p_n/q_n = (p_0q_0)\,(s_0s_1s_2 \cdot \cdot \cdot s_{n-1})$$

If this is written as $p_n/q_n = (p_0/q_0)s^n$, then how can s be interpreted?

14 If selection is not too strong, the time required for the allele frequency of a favored dominant allele to change from p_0 to p_t is given by

$$\ln(p_t/q_t) + (1/q_t) = [\ln(p_0/q_0) + (1/q_0)] + st$$

Use this to derive the analogous equation for a favored recessive.

15 The following equation has equilibria at $p = 0$, 1/2, and 1. Classify them as to stability. If there is a stable equilibrium, is it locally or globally stable?

$$\Delta p = p(1/2 - p)(1 - p)$$

16 Suppose the viabilities of males and females are as follows:

	AA	*Aa*	*aa*
Females	0.9	1	0.8
Males	1.0	v_{12}	0.5

What is the smallest value of v_{12} that ensures a protected polymorphism?

17 If allele a is a recessive lethal in zygotes and the relative fitness of A:a gametes is $1 - s$:1, then what is the equilibrium allele frequency of a?

18 In a *Drosophila* population cage containing a meiotically driven chromosome known as *segregation distorter*, the equilibrium frequency of the driven chromosome was approximately 0.125 and the segregation ratio in heterozygotes was about $k = 0.75$ (Hiraizumi et al. 1960). The meiotic drive chromosome is homozygous lethal in both sexes. Use Equation 4.66 to estimate the approximate value of w_{12} consistent with these data.

19 Show that the allele frequency of a recessive lethal is given by $q_n = q_0/(1 + nq_0)$. (Hint: It is easiest to derive an expression first for $1/q_n$.) How many generations are required to reduce the allele frequency by half?

20 For an allele A at frequency p_0 with AA, Aa, and aa, fitnesses of $1 + s$, $1 + s/2$, and 1, the analogue of Equation 4.74 for the ultimate probability of fixation is $\text{Prob}(A \text{ fixes}) = [1 - \exp(-2Nsp_0)]/[1 - \exp(-2Ns)]$. Calculate for $p_0 = 0.1, 0.2, 0.3, \ldots 0.9$, given $Ns = 0.1$ and $Ns = 10$.

21 Equation 4.74 for the probability of ultimate fixation of a single copy of a mutant gene is often given in the form $\text{Prob}(A \text{ fixes}) = 2s/[1 - \exp(-4Ns)]$ (e.g., Crow and Kimura 1970). This applies to a model with fitnesses $1 + 2s$, $1 + s$, 1. Show that this formulation is equivalent to Equation 4.74.

22 Use the following definitions: $w_1 = pw_{11} + qw_{12}$; $w_2 = pw_{12} + qw_{22}$; $V_g = 2p(w_1 - \overline{w})^2 + 2q(w_2 - \overline{w})^2$; and $\alpha = p(w_{11} - w_{12}) + q(w_{12} - w_{22})$ to show that

a. $\alpha = w_1 - w_2$

b. $\overline{w} = pw_1 + qw_2$

c. $V_g = 2pq\alpha^2$

d. $\Delta p = pq\alpha / \overline{w}$

23 Expand $\overline{w}' = w_{11}(p + \Delta p)^2 + 2w_{12}(p + \Delta p)(q - \Delta p) + w_{22}(q - \Delta p)^2$ and use the results in Problem 22 to show that

$$\Delta\overline{w} = (V_g/\overline{w})[1 + (pq/2\overline{w})(w_{11} - 2w_{12} + w_{22})]$$

This is the exact result for the discrete model. The proof in the text assumes $dp/dt \approx \Delta p$, which eliminates the second term in the brackets. The approximation is best for small differences in fitness.

Chapter Five

Inbreeding and Other Forms of Nonrandom Mating

Inbreeding refers to matings between related individuals, and the primary effect of inbreeding is a departure from Hardy-Weinberg genotype frequencies toward an excess of homozygotes. In this chapter we show how to quantify the amount of inbreeding in a population, and formal models are explored that consider the long-term evolutionary effects of inbreeding. In the absence of any other evolutionary factors, *inbreeding alone does not change allele frequencies*—it changes only the assembly of genes into genotypes. However, if natural selection is also operating, then inbreeding can have a profound effect on the course of evolution. Similarly, inbreeding and random genetic drift are overlapping phenomena, because random genetic drift results in an increased likelihood of bringing together genes from a common ancestor due to loss of genetic variation.

In populations with a low allele frequency of deleterious recessives, such as humans, inbreeding can result in a much higher frequency of affected genotypes, since the frequency of homozygotes for the deleterious recessive alleles is increased. Because rare recessives can hide in heterozygous individuals, inbreeding has the paradoxical beneficial effect of exposing the deleterious alleles and thereby increasing the power of natural selection to remove them. Inbreeding, unlike assortative mating, affects all genes in inbred individuals. The most extreme example of inbreeding is self-fertilization, common in grasses, in which the population is essentially an assemblage of separate clones. In organisms that are not so completely inbred, it is necessary to quantify the amount of inbreeding in a way that is easy to estimate and is useful theoretically. This is the subject of the next section.

THE INBREEDING COEFFICIENT

The inbreeding coefficient F has an interpretation in terms of probability that is extremely useful for calculation of F from pedigrees. To express the inbreeding coefficient in terms of probability, imagine the two alleles present in homologous chromosomes in a single inbred individual. These two alleles could be identical by descent — that is, they could both have been derived by replication of a single allele in a common ancestor. If the two alleles are identical by descent, they are **autozygous**. On the other hand, the alleles may not be replicas of a single ancestral allele, in which case the alleles are not identical by descent, and are **allozygous**. The interpretation of F in terms of probability is this: *The inbreeding coefficient F is the probability that the two alleles of a gene in an individual are identical by descent (autozygous).*

Seen in terms of the probability of autozygosity, the inbreeding coefficient is clearly a relative concept in that it measures the amount of autozygosity in the present population relative to that in some ancestral population. It can be arbitrarily assumed that all alleles in the ancestral population are not identical by descent, and the inbreeding coefficient of an individual in the present population is then the probability that the two homologous alleles were derived by replication of a common ancestral allele. The ancestral population need not be remote in time from the present population. Indeed, the ancestral population, presumed to be noninbred ($F = 0$), typically refers to the population existing just a few generations prior to the present one, and F in the present population then measures inbreeding that has occurred in the span of these few generations. Because the span of time involved is usually short, the possibility of mutation can usually be ignored. Autozygous individuals must therefore be homozygous for some allele at the locus under consideration. On the other hand, allozygous individuals can be either homozygous or heterozygous.

Even though the inbreeding coefficient is defined as the probability of an individual having a pair of alleles that are identical by descent, there are two uses of the inbreeding coefficient that are subtly distinct. When the inbreeding coefficient is used in models, such as models that predict genotype frequencies, then it is a **parameter**, and is assumed to be known without error. On the other hand, when F is estimated from population heterozygosity or from a pedigree, then it is a statistical **estimator**, and as such represents a description of an observation. In Chapter 3 we used the term *fixation index* to refer to F as a measure of departure from panmictic proportions (for two alleles, $F = 1 - H_{obs}/2pq$, where H_{obs} is the observed heterozygosity). Often these two forms are expressed by different symbols, such as f for the parameter and F for the estimator. We will use the symbol F for

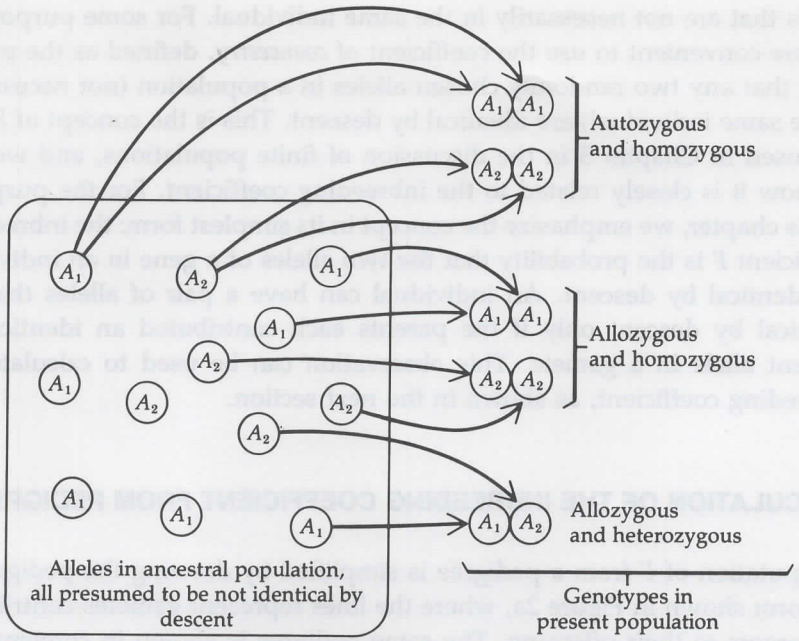


Figure 1. In an autozygous individual, homologous alleles are derived from replication of a single DNA sequence in an ancestor, and they are therefore identical by descent. In an allozygous individual, homologous alleles are not identical by descent. As shown here, allozygous individuals may be heterozygous or homozygous, but autozygous individuals must be homozygous (except in the unlikely event that one allele undergoes a mutation).

both meanings to avoid the "polymorphism" of f and F in the literature, hoping the meaning in each case will be clear from context.

Figure 1 illustrates how the concepts of autozygosity and allozygosity are related to those of homozygosity and heterozygosity. The essential point is that two alleles can be chemically identical (in terms of having the same sequence of nucleotides along the DNA) without being identical by descent; the concept of identity by descent pertains to the ancestral origin of an allele and not to its chemical makeup. Although, as shown in Figure 1, two distinct alleles that are chemically identical (two A's or two a's, for example) may come together in an individual and thereby make the individual homozygous, the alleles in the ancestral population are, by definition, not identical by descent, so the individual is allozygous. Similarly, although a heterozygous individual must be allozygous (ignoring mutation), a homozygous individual may be either autozygous or allozygous.

Although the inbreeding coefficient F is the probability that the two alleles in an individual are identical by descent, the idea can readily be extended to

alleles that are not necessarily in the same individual. For some purposes it is more convenient to use the coefficient of *coancestry*, defined as the probability that any two randomly chosen alleles in a population (not necessarily in the same individual) are identical by descent. This is the concept of F that was used in Chapter 3 in the discussion of finite populations, and we will see how it is closely related to the inbreeding coefficient. For the purposes of this chapter, we emphasize the concept in its simplest form: the inbreeding coefficient F is the probability that the two alleles of a gene in an individual are identical by descent. An individual can have a pair of alleles that are identical by descent only if the parents each contributed an identical by descent allele in a gamete. This observation can be used to calculate the inbreeding coefficient, as shown in the next section.

CALCULATION OF THE INBREEDING COEFFICIENT FROM PEDIGREES

Computation of F from a pedigree is simplified by drawing the pedigree in the form shown in Figure 2a, where the lines represent gametes contributed by parents to their offspring. The same pedigree is shown in conventional form in Figure 2b. The individuals in gray in panel (b) are not represented in panel (a) because they have no ancestors in common and therefore do not contribute to the inbreeding of individual I. The inbreeding coefficient F_I of individual I is the probability that I is autozygous for alleles of the autosomal gene under consideration. This is the same as F_{DE}, which is the probability

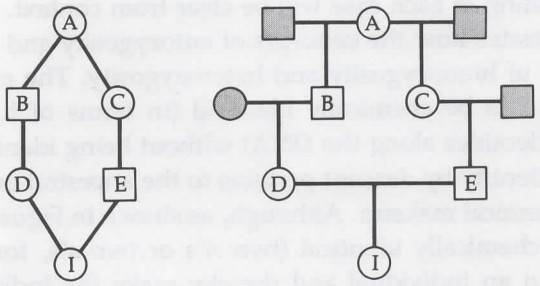


Figure 2. (a) A convenient way to represent pedigrees for calculation of the inbreeding coefficient. In this case, the pedigree shows a mating between half-first cousins. (b) Conventional representation of the same pedigree as in panel a. Squares represent males, circles represent females, and the shaded individuals in panel b are not depicted in panel a because they do not contribute to the inbreeding of the individual designated I.

that gametes D and E are identical. The chance that one allele from one individual is identical by descent to an allele drawn from another individual is the coefficient of **coancestry**. The coefficient of inbreeding is therefore equal to the coefficient of coancestry of the individual's parents.

The first step in calculating F is to locate all the common ancestors in the pedigree, because an allele can become autozygous in I only if it was inherited through both of I's parents from a common ancestor; in this case, there is only one common ancestor — the individual labeled A. The next step in calculating F, which is carried out for each common ancestor, is to trace all the paths of gametes that lead from one of I's parents back to the common ancestor and then down again to the other parent of I. These paths are the paths along which an allele in a common ancestor could become autozygous in the individual I. In Figure 2a, there is only one such path: DBACE, (the common ancestor is underlined for bookkeeping purposes, an especially useful procedure in complex pedigrees).

The third step in calculating F is to calculate the probability of autozygosity in I due to each of the paths in turn. For the path DBACE, the reasoning involved is illustrated in Figure 3. Here the black dots represent alleles transmitted along the gametic paths, and the number associated with each loop is the probability of identity by descent of the alleles indicated. For all individuals except the common ancestor, the probability is 1/2, because with Mendelian segregation the probability that a particular allele present in a parent is transmitted to a specified offspring is 1/2. To see why $(1/2)(1 + F_A)$ is the probability associated with the loop around the common ancestor,

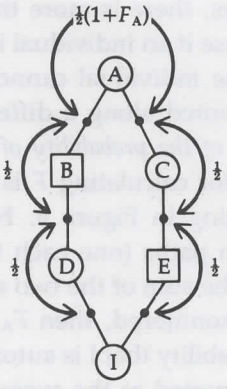


Figure 3. Loops for the pedigree in Figure 2a, showing probabilities that designated alleles (solid dots) are identical by descent. Each loop is independent of the others, so their probabilities multiply. Thus, the inbreeding coefficient of individual I is $F_I = (1/2)^5(1 + F_A)$, where F_A represents the inbreeding coefficient of the common ancestor.

denote the alleles in the common ancestor as α_1 and α_2. (These symbols are used to avoid confusion with conventional allele symbols designating chemical types of alleles, such as *A* and *a* for dominant and recessive, respectively.) The pair of gametes contributed by individual *A* could contain $\alpha_1\alpha_1$, $\alpha_2\alpha_2$, $\alpha_1\alpha_2$, or $\alpha_2\alpha_1$, each with probability 1/4 because of Mendelian segregation. In the first two cases, the alleles are clearly identical by descent; in the second two cases, the alleles are identical by descent only if α_1 and α_2 are identical by descent, and α_1 and α_2 are identical by descent only if individual A is autozygous, which has probability F_A (the inbreeding coefficient of A). Altogether, the required probability for the loop around individual A is

$$1/4 + 1/4 + (1/4)F_A + (1/4)F_A = 1/2 + (1/2)F_A = (1/2)(1 + F_A) \quad \textbf{(5.1)}$$

Now, each of the steps in Figure 3 is independent of the others, so the total probability of autozygosity in individual I due to this path is

$$1/2 \times 1/2 \times (1/2)(1 + F_A) \times 1/2 \times 1/2$$

or

$$(1/2)^5(1 + F_A) \quad \textbf{(5.2)}$$

Note that the exponent on the 1/2 is simply the number of individuals in the path. In general, if a path through a common ancestor A contains *i* individuals, the probability of autozygosity due to that path is

$$(1/2)^i(1 + F_A) \quad \textbf{(5.3)}$$

Thus, the inbreeding coefficient of individual I in Figure 2a is $(1/2)^5(1 + F_A)$, or, assuming $F_A = 0$, simply $(1/2)^5 = 1/32$.

In more complex pedigrees, there is more than a single path. The paths are mutually exclusive, because if an individual is autozygous due to an allele inherited along one path, the individual cannot at the same time be autozygous due to an allele inherited along a different path. *Thus, the total inbreeding coefficient is the sum of the probability of autozygosity for each separate path.* The whole procedure for calculating *F* is summarized in an example involving a first cousin mating in Figure 4. Here there are two common ancestors (A and B) and two paths (one each through A and B). The total inbreeding coefficient of I is the sum of the two separate contributions shown in Figure 4. If A and B are noninbred, then $F_A = F_B = 0$, and $F_I = (1/2)^5 + (1/2)^5 = 1/16$. This is the probability that I is autozygous at the specified locus. Alternatively, F_I can be interpreted as the average proportion of all genes in I in which the alleles are autozygous.

In general, for any autosomal gene, the formula for calculating the inbreeding coefficient F_I of an individual I is

$$F_I = \Sigma\,(1/2)^i(1 + F_A) \quad \textbf{(5.4)}$$

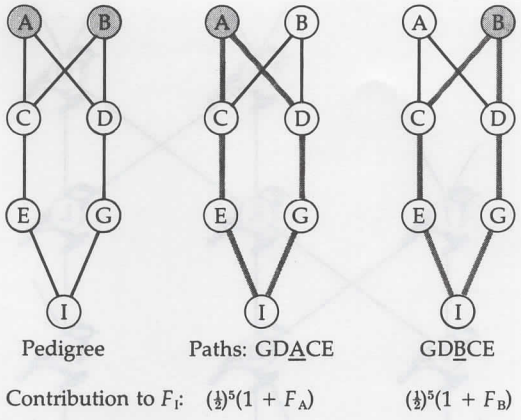


Figure 4. On the left is a pedigree of an individual I, the child of a first cousin mating. On the right are the two paths through common ancestors (heavy lines) involved in calculating the inbreeding coefficient of I. Below each path is the contribution to F_1 due to that path, calculated as in Figure 3. Each path is mutually exclusive of the others, so their probabilities add. Thus, the total inbreeding coefficient of I is the sum of the two separate contributions.

where the summation is over all possible paths through all common ancestors, i is the number of individuals in each path, and A is the common ancestor in each path. The method for calculating F that works for arbitrarily complex pedigrees is presented in Box A.

Inbreeding in Irregular Pedigrees

A

Figure 5 depicts a pedigree of great tits on the Dutch island of Vlieland, constructed by painstaking banding and following of birds over a period of years (van Noordwijk and Scharloo 1981). In order to apply Equation 5.4 to this pedigree, it is necessary first to tabulate the possible paths of coancestry.

Common ancestor of A and B	Path	Path length	F_A	$(1/2)^i(1 + F_A)$
F	BCFA	4	0	$(1/2)^4$
H	BHEA	4	0	$(1/2)^4$
K	BCDIKHEA	8	0	$(1/2)^8$
J	BCDIJHEA	8	0	$(1/2)^8$
M	BHKMLGFA	8	0	$(1/2)^8$
N	BHKNLGFA	8	0	$(1/2)^8$
N	BCDIKNLGFA	10	0	$(1/2)^{10}$
M	BCDIKMLGFA	10	0	$(1/2)^{10}$
M	BCFGLMKHEA	10	0	$(1/2)^{10}$
N	BCFGLNKHEA	10	0	$(1/2)^{10}$

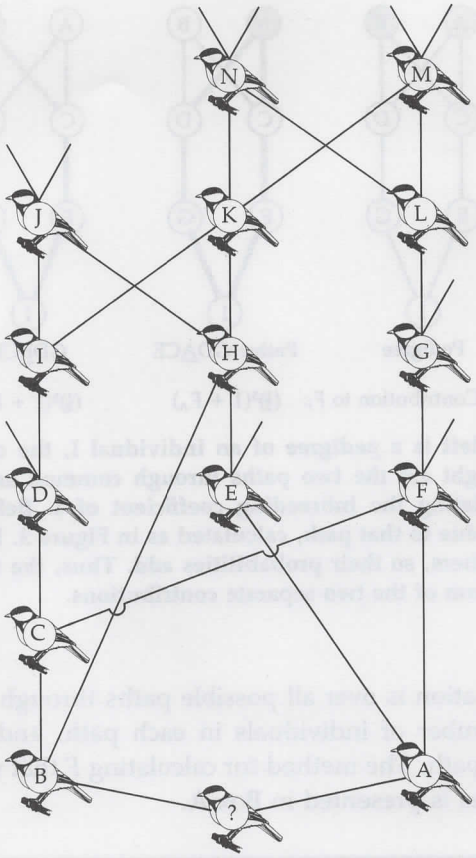


Figure 5. A complicated pedigree of inbred great tits from Vlieland. All individuals shown are at least related as great-grandparents to A or B. As shown in the text, there are ten relations between A and B ($F = 0.145$). (From van Noordwijk and Scharloo 1981.)

The coefficient of inbreeding of the bottom individual in the pedigree is the sum of the right hand column, or $F = 0.145$.

For pedigrees much more complex than this it is desirable to resort to a computer to tally the paths of coancestry. An application of intensive computer searching of pedigrees was done with standardbred and thoroughbred horses (MacCluer et al. 1983). Thirty generations of data had been kept on over 5000 horses, and programs were written to calculate F for the various members of the pedigree. With deeper pedigrees that go further back in time, there are more opportunities for paths of coancestry, as the above example illustrates. In the horse pedigrees, with 30 generations of data, the number of paths of coancestry became staggering. For example, the average horse born in 1980 had 218,202 paths of coancestry, and the horse with the greatest number of paths had 1,013,352! Since breeders try to avoid close inbreeding, the paths are typically fairly long. The average path length in the horses was 6.13 individuals. Despite the multiplicity of paths, each contributes very little to the inbreeding coefficient because

most of them are quite long. The overall levels of inbreeding in horses in the years 1900, 1940, and 1980 were 0.0077, 0.0479, and 0.0899, respectively. It is important to note that if the analysis had been done using only the last four or five generations, the value of the inbreeding coefficient would have been seriously underestimated.

The procedure for calculating F in Equation 5.4 is valid only for autosomal (i.e., non-X-linked) genes. An analogous sort of procedure does exist for calculating F for X-linked genes, as outlined in Box B.

Inbreeding Coefficient for X-Linked Genes

B

The X chromosome presents some interesting twists to the calculation of the inbreeding coefficient. First of all, normal males have only one copy of the X chromosome and can never have a pair of X-linked alleles identical by descent, so F for X-linked genes in males is always 0. Figure 6 presents the possible paths of coancestry for X-linked alleles in a daughter of first cousins. There are four configurations for the first cousins: their fathers may be siblings, their mothers may be siblings, or either of two ways in which one mother and one father are siblings. In the first case, since males transmit only the Y chromosome to sons, there is no complete path of coancestry, so daughters of first cousin males are not inbred with respect to the X chromosome. If the cousins have mothers who are siblings, then the configuration is precisely like an autosomal gene, but the coefficient of inbreeding is not the same, because any male in a path must transmit the same X chromosome he received from his mother. Since loops around males do not contribute the factor 1/2 to F, the simplest adjustment to calculate F is to count only females in the path. The second diagram in Figure 6 has one path with three females and two males, and another path with four females and one male, so $F = (1/2)^3 + (1/2)^4 = 3/16$. The third type of first cousins has no path of coancestry, because

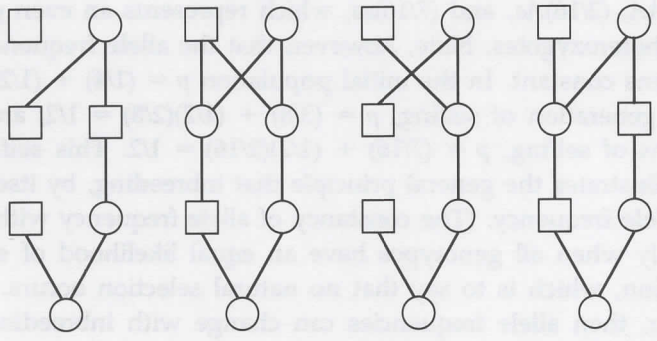


Figure 6. Calculation of F for X-linked genes. There are four possible configurations that must be considered for first cousins.

an X chromosome cannot be transmitted through two successive males. The fourth type of first cousin has an inbreeding coefficient of 1/8, since there are three females in the path through the female in the first generation.

The rules for calculating the inbreeding coefficient for X-linked genes can be summarized as (1) ignore all paths with two or more consecutive males; and (2) ignore males when counting the path lengths. Other than these two rules, Equation 5.4 applies to arbitrarily complex pedigrees for X-linked genes.

GENOTYPE FREQUENCIES WITH INBREEDING

The principal effect of inbreeding in a population is to increase the frequency of homozygous genotypes at the expense of the frequency of heterozygous genotypes. This effect can most easily be seen in the case of repeated self-fertilization. Consider a self-fertilizing population of plants that consists of (1/4)*AA*, (1/2)*Aa*, and (1/4)*aa* genotypes. These genotype frequencies are in Hardy-Weinberg proportions. Because each plant undergoes self-fertilization, the *AA* and *aa* genotypes produce only *AA* and *aa* offspring, respectively, and the *Aa* genotypes produce (1/4)*AA*, (1/2)*Aa*, and (1/4)*aa* offspring. After one generation of self-fertilization, therefore, the genotype frequencies of *AA*, *Aa*, and *aa* are

$$\begin{array}{ll} AA & (1/4)(1) + (1/2)(1/4) \\ Aa & (1/2)(1/2) \\ aa & (1/4)(1) + (1/2)(1/4) \end{array}$$

or (3/8)*AA*, (2/8)*Aa*, and (3/8)*aa* (see Figure 7). These genotype frequencies are no longer in the Hardy-Weinberg proportions because there is a deficiency of heterozygous genotypes and an excess of homozygous genotypes.

After a second generation of self-fertilization, the genotype frequencies are (7/16)*AA*, (2/16)*Aa*, and (7/16)*aa*, which represents an even greater deficiency of heterozygotes. Note, however, that the allele frequency of *A* (call it p) remains constant. In the initial population $p = (1/4) + (1/2)(1/2) = 1/2$; after one generation of selfing, $p = (3/8) + (1/2)(2/8) = 1/2$; and after two generations of selfing, $p = (7/16) + (1/2)(2/16) = 1/2$. This self-fertilization example illustrates the general principle that inbreeding, by itself, does not change allele frequency. (The constancy of allele frequency with inbreeding occurs only when all genotypes have an equal likelihood of survival and reproduction, which is to say that no natural selection occurs. If selection does occur, then allele frequencies can change with inbreeding, but in a different manner than with random mating.)

The effects of inbreeding are conveniently measured in terms of the re-

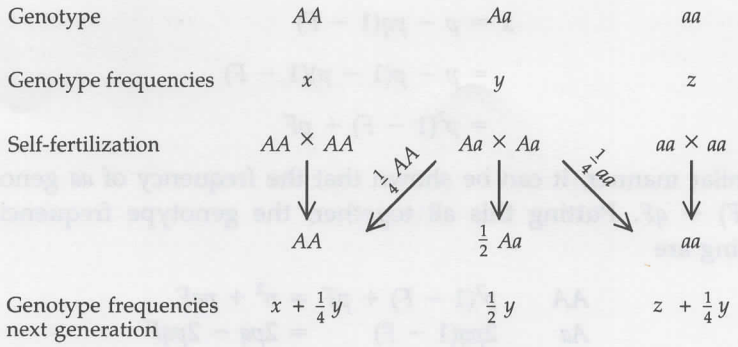


Figure 7. Changes in genotype frequencies under self-fertilization. When homozygotes self-fertilize they produce progeny of their own genotype, but when heterozygotes self-fertilize only half of their offspring are heterozygotes. Thus selfing results in a halving of heterozygosity each generation.

duction in heterozygosity that occurs. That is, we can measure the amount of inbreeding by comparing the actual proportion of heterozygous genotypes in the population with the proportion of heterozygous genotypes that would occur with random mating. To be precise, consider a gene with two alleles, A and a, at respective frequencies p and q (with $p + q = 1$). Suppose the actual frequency of heterozygous genotypes in the population is H. If the population were undergoing random mating for the gene, the frequency of heterozygous genotypes would be $2pq$ because of the Hardy-Weinberg principle, but for the sake of generality we will denote the random-mating heterozygosity by the symbol H_0. The effects of inbreeding can be defined in terms of the quantity $(H_0 - H)/H_0$, which, as will be shown later, is another way to write the inbreeding coefficient F. That is,

$$F = (H_0 - H)/H_0 \tag{5.5}$$

In biological terms, F measures the fractional reduction in heterozygosity, relative to a random-mating population with the same allele frequencies. Because $H_0 = 2pq$, the actual frequency of heterozygous genotypes in the inbred population can be written in terms of F as $H = H_0 - H_0F = H_0(1 - F) = 2pq(1 - F)$.

The frequency of AA homozygous genotypes in the inbred population can also be expressed in terms of F. Suppose that the proportion of AA genotypes is actually x. Because the allele frequency of A is p, we must have, $x + H/2 = p$. But $H = 2pq(1 - F)$, thus

$$x = p - 2pq(1 - F)/2 \tag{5.6}$$

Some algebra puts this into another form that is usually more useful:

$$
\begin{aligned}
x &= p - pq(1 - F) \\
&= p - p(1 - p)(1 - F) \\
&= p^2(1 - F) + pF
\end{aligned}
\tag{5.7}
$$

In a similar manner, it can be shown that the frequency of *aa* genotypes is $q^2(1 - F) + qF$. Putting this all together, the genotype frequencies with inbreeding are

$$
\begin{array}{lll}
AA & p^2(1 - F) + pF & = p^2 + pqF \\
Aa & 2pq(1 - F) & = 2pq - 2pqF \\
aa & q^2(1 - F) + qF & = q^2 + pqF
\end{array}
\tag{5.8}
$$

These expressions were first derived by Wright (1921). The left hand expressions for the genotype frequencies can be partitioned into a component of size $(1 - F)$ as with random mating and a component of size F with complete inbreeding (when the genotype frequencies are p, 0, q). Symbolically, the genotype frequencies of *AA*, *Aa*, and *aa* can be obtained by multiplying:

$$
(1 - F)(p^2, 2pq, q^2) + F(p, 0, q) \tag{5.9}
$$

Expression 5.9 facilitates comparison with the genotype frequencies in the Hardy-Weinberg principle. Notice that, when $F = 0$, which represents no inbreeding (i.e., random mating), the genotype frequencies are in the familiar Hardy-Weinberg proportions p^2, $2pq$, and q^2. When $F = 1$, which represents complete inbreeding, the population consists entirely of *AA* and *aa* homozygotes in the frequencies p and q, respectively. Another way to represent the genotype frequencies that is also equivalent to Equation 5.8 is

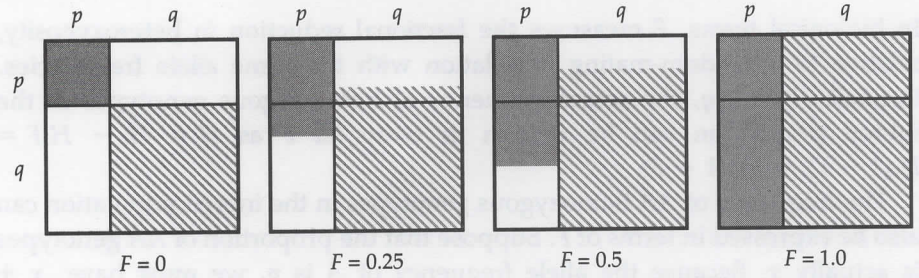


Figure 8. Genotype frequencies in an inbred population can be represented on these modified Punnett squares, where $p \neq q$. The gray area represents the frequency of *AA* and the hatched area represents the frequency of *aa*. With $F = 0$, the white areas represent a heterozygote frequency of $2pq$. Note that the fraction of the heterozygote rectangles that is filled in by homozygotes is F.

AA	$p - (1 - F)pq$
Aa	$(1 - F)2pq$
aa	$q - (1 - F)pq$

In this case, the genotype frequencies are expressed as deviations from the case of $F = 1$, showing that the frequency of genotype AA decreases in the amount $(1 - F)pq$ as F decreases from the value of 1. Figure 8 is a graphical representation of the situation, and Figure 9 is a De Finetti diagram showing

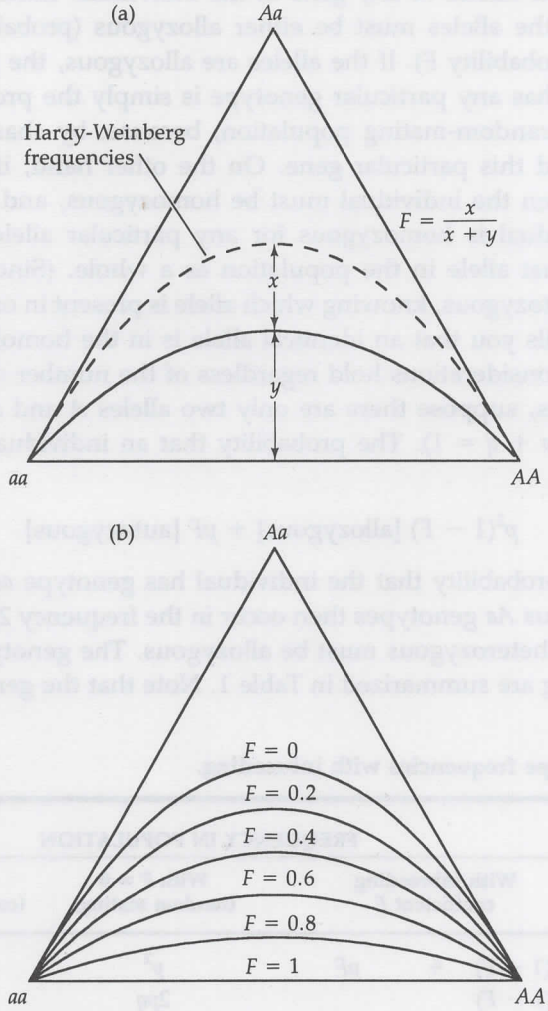


Figure 9. (a) The genotype frequencies in an inbred population correspond to a parabola on the De Finetti diagram that lies below the Hardy-Weinberg parabola. The fractional decrease in the frequency of heterozygotes is *F*. (b) Genotype frequencies for several values of *F*.

the frequencies of genotypes *AA*, *Aa*, and *aa* under several different levels of inbreeding. It is clear from Figure 9 that F is the fractional reduction in heterozygote frequency compared to the panmictic or Hardy-Weinberg expected frequency.

To show that the inbreeding coefficient defined in Equation 5.5 is equivalent to the probability of autozygosity, we need only consider the implications of the probability definition for an entire population. For this purpose, imagine a population with inbreeding coefficient F. Focus on one individual, and consider the alleles of any gene in the individual. Either of two things must be true: the alleles must be either allozygous (probability $1 - F$) or autozygous (probability F). If the alleles are allozygous, the probability that the individual has any particular genotype is simply the probability of that genotype in a random-mating population, because, by chance, inbreeding has not affected this particular gene. On the other hand, if the alleles are autozygous, then the individual must be homozygous, and the probability that the individual is homozygous for any particular allele is simply the frequency of that allele in the population as a whole. (Since the alleles in question are autozygous, knowing which allele is present in one chromosome immediately tells you that an identical allele is in the homologous chromosome.) These considerations hold regardless of the number of alleles, but to simplify matters, suppose there are only two alleles A and a at frequencies p and q (with $p + q = 1$). The probability that an individual has genotype *AA* is therefore

$$p^2(1 - F) \text{ [allozygous]} + pF \text{ [autozygous]}$$

Similarly, the probability that the individual has genotype *aa* is $q^2(1 - F) + qF$. Heterozygous *Aa* genotypes then occur in the frequency $2pq(1 - F)$, since alleles that are heterozygous must be allozygous. The genotype frequencies with inbreeding are summarized in Table 1. Note that the genotype frequen-

Table 1. Genotype frequencies with inbreeding.

Genotype	FREQUENCY IN POPULATION			
	With inbreeding coefficient F		With $F = 0$ (random mating)	With $F = 1$ (complete inbreeding)
AA	$p^2(1 - F)$ +	pF	p^2	p
Aa	$2pq(1 - F)$		$2pq$	0
aa	$q^2(1 - F)$ +	qF	q^2	q
	Allozygous genes	Autozygous genes		

cies are exactly the same as those given in Equation 5.8, which shows that the autozygosity definition of F and the heterozygosity definition of F, though superficially quite different, are actually equivalent. Box C shows yet another interpretation for the inbreeding coefficient, which was the original one as defined by Wright.

C

F is the Correlation of Uniting Gametes

To show that the inbreeding coefficient F is the same as the correlation between uniting gametes, assign genetic values of 1 to allele A and 0 to allele a. We then construct a table showing the frequencies of various zygotes and calculate the correlation coefficient of these genetic values.

Egg	Sperm	Zygote frequency	Genetic value: Egg (x)	Genetic value: Sperm (y)	x^2	y^2	xy
A	A	$p^2(1-F)+pF$	1	1	1	1	1
A	a	$pq(1-F)$	1	0	1	0	0
a	A	$pq(1-F)$	0	1	0	1	0
a	a	$q^2(1-F)+qF$	0	0	0	0	0

The definition of the correlation coefficient ρ is

$$\rho = \frac{\text{cov}(x,y)}{\sqrt{\text{Var}(x)\text{Var}(y)}} = \frac{\overline{xy} - \bar{x}\bar{y}}{\sqrt{(\overline{x^2} - \bar{x}^2)(\overline{y^2} - \bar{y}^2)}}$$

The mean genetic value of eggs is

$$\bar{x} = p^2(1-F) + pF + pq(1-F) = p$$

This is also the mean genetic value of sperm. Similarly, the mean of both the x^2 and the y^2 columns is p. The mean of the product, $\overline{xy}$ is $p^2(1-F) + pF$. Substituting into the formula for the correlation coefficient, we get

$$\rho = \frac{p^2(1-F) + pF - p^2}{\sqrt{(p-p^2)(p-p^2)}} = \frac{-Fp^2 + Fp}{p - p^2} = F$$

This completes the proof that the inbreeding coefficient is the correlation coefficient of uniting gametes. With random mating, gametes unite at random, and there is no correlation of genetic values of uniting gametes. Hence $\rho = 0$ and $F = 0$. With total inbreeding, $F = 1$ and only gametes that are identical by descent unite, so they are absolutely correlated and $\rho = 1$. For intermediate values, the equivalence of F and the correlation of uniting gametes still holds. Wright always preferred the definition of F as the correlation between uniting gametes (rather than probability of autozygosity), because with the correlation definition (but not with the probability definition) F can be negative.

If a population contains multiple alleles A_1, A_2, . . . , A_n at respective frequencies p_1, p_2, . . . , p_n (with $p_1 + p_2 + \ldots + p_n = 1$), then with inbreeding coefficient F, the frequencies of A_iA_i homozygotes and A_iA_j heterozygotes are

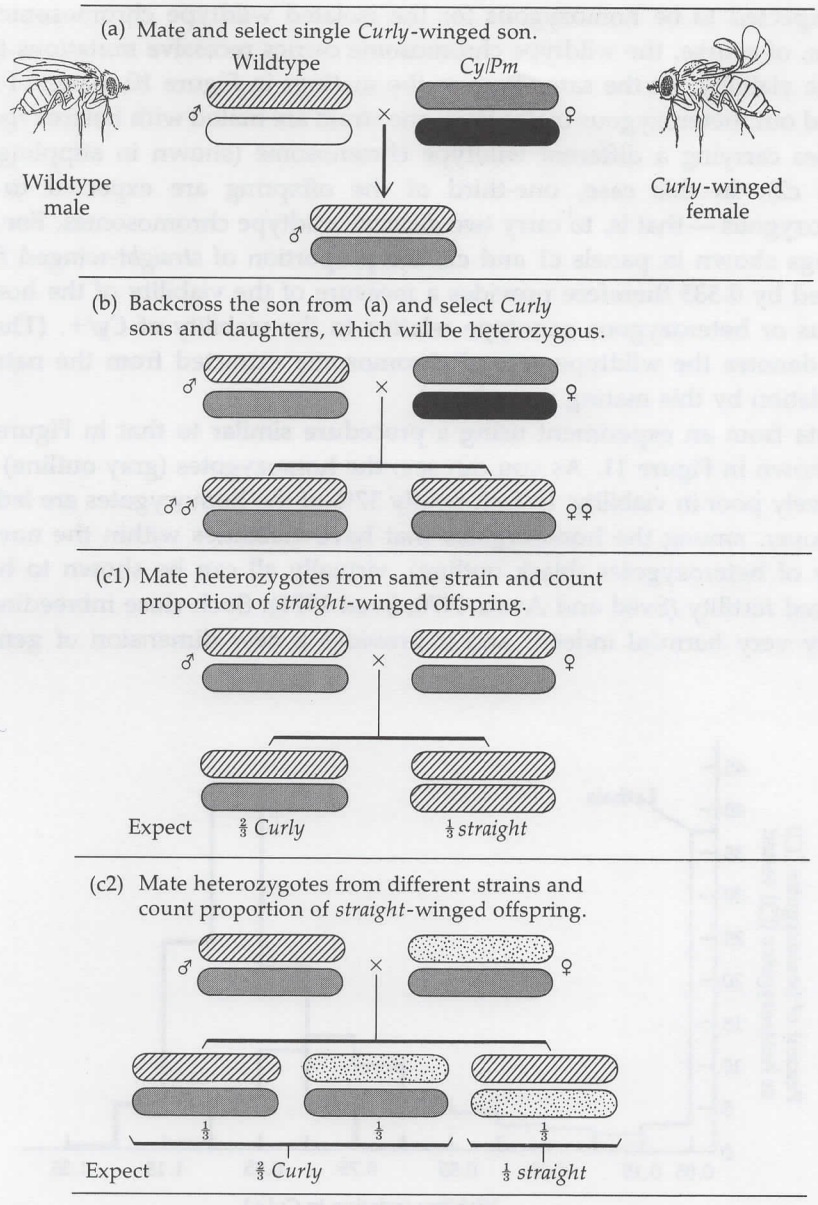
$$\begin{array}{ll} A_iA_i \text{ homozygotes} & p_i^2(1 - F) + p_iF \\ A_iA_j \text{ heterozygotes} & 2p_ip_j(1 - F) \end{array} \qquad (5.11)$$

Thus inbreeding causes a universal increase in the frequency of homozygotes above the Hardy-Weinberg proportions, and a corresponding decrease in the frequency of heterozygotes. One important consequence is seen in the inflated incidence of human recessive disorders among offspring from consanguineous matings.

Effects of Inbreeding and the Frequency of Human Genetic Disorders

In **outcrossing** species — those that regularly avoid inbreeding — close inbreeding is generally harmful. The effects are seen most dramatically when inbreeding is complete or nearly complete. Although nearly complete autozygosity can be approached in most species by many generations of brother–sister mating, autozygosity of individual chromosomes can easily be accomplished in *Drosophila* by the sort of mating scheme shown in Figure 10. This design is very similar to the approach used by Mukai to examine the effects of spontaneous mutations, discussed in Chapter 3. *Cy* (*Curly* wings) and *Pm* (*Plum*-colored eyes) are dominant mutations on chromosome 2 and are present on certain laboratory second chromosomes (denoted by gray and black) that carry several long inversions to prevent recombination. Single wildtype males are mated individually to *Cy*/*Pm* females, and a single *Curly*-winged son, carrying one wildtype second chromosome (hatched), is selected from each mating (Figure 10a). The *Curly*-winged male is backcrossed to *Cy*/*Pm* females, and *Curly*-winged heterozygous sons and daughters are selected (Figure 10b). These are mated with each other and the offspring counted. Because *Cy*/*Cy* homozygotes are lethal, one-third of the surviving offspring

Figure 10. A mating scheme to extract single chromosomes (in this case, the second chromosome) from populations of *Drosophila melanogaster*. *Cy* (*Curly* wings) and *Pm* (*Plum* eye color) are dominant mutations contained in certain special laboratory strains of flies that contain chromosomes that have multiple inversions to prevent recombination. From each mating of the type in panel a, a single son (containing one of his father's second chromosomes, indicated by the hatching) is selected. This son is backcrossed (panel b) in order to produce many replicas of that second chromosome. Brother-sister matings as in panel c1 are expected to produce 1/4 *Cy*/*Cy*, 1/2 *Cy*/+, and 1/4 +/+ zygotes (the + denotes the wildtype second chromosome),



but since *Cy*/*Cy* zygotes do not survive, the expected offspring are 2/3 *Cy*/+ (*Curly* wings) and 1/3 +/+ (normal, *straight* wings). Matings as in panel c2, between a male containing one wildtype second chromosome (hatched) and a female carrying a different one (stippled), are also expected to produce 2/3 *Curly* wings and 1/3 *straight* wings. However, in panel c1, the *straight*-wing flies are homozygous for a single wildtype second chromosome, so deleterious recessive alleles will result in a lower count of the *straight*-wing flies. In panel c2 the *straight*-wing flies are heterozygous for two different wildtype second chromosomes.

are expected to be homozygous for the isolated wildtype chromosome — unless, of course, the wildtype chromosome carries recessive mutations that reduce viability. At the same time as the matings in Figure 10, panel c1 are carried out, heterozygous males from one strain are mated with heterozygous females carrying a different wildtype chromosome (shown in stippling in panel c2). In this case, one-third of the offspring are expected to be heterozygous — that is, to carry two *different* wildtype chromosomes. For the matings shown in panels c1 and c2, the proportion of *straight*-winged flies divided by 0.333 therefore provides a measure of the viability of the homozygous or heterozygous genotype relative to the viability of *Cy*/+. (The + here denotes the wildtype second chromosome extracted from the natural population by this mating procedure.)

Data from an experiment using a procedure similar to that in Figure 10 are shown in Figure 11. As you can see, the homozygotes (gray outline) are relatively poor in viability. In fact, nearly 37% of the homozygotes are lethal. Moreover, among the homozygotes that have viabilities within the normal range of heterozygotes (black outline), virtually all can be shown to have reduced fertility (Sved and Ayala 1970; Sved 1971). Such close inbreeding is clearly very harmful indeed, and it provides a new dimension of genetic

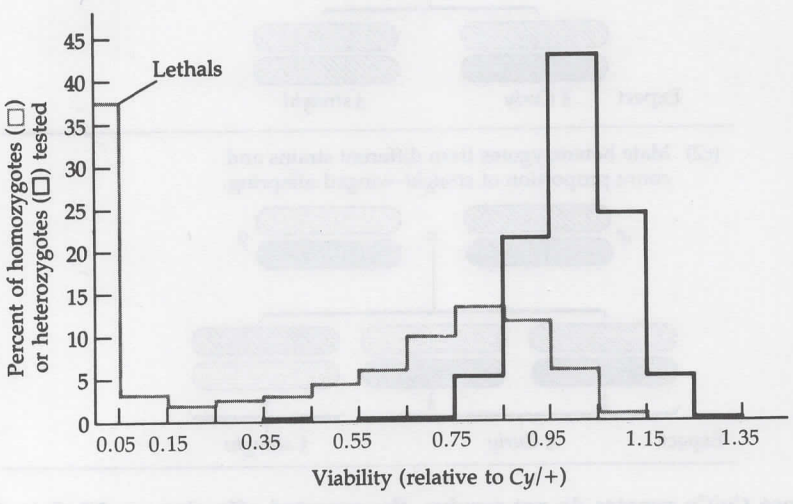


Figure 11. Viability distributions of wildtype homozygotes (gray outline) and wildtype heterozygotes (black outline) of second chromosomes extracted from *Drosophila melanogaster* according to the mating scheme in Figure 10. The histograms depict results of testing 691 homozygous combinations and 688 heterozygous combinations. Note that, in this sample, nearly 37% of the wild chromosomes are homozygous lethal, and many more have viabilities substantially below normal. (Data from Mukai and Yamaguchi 1974.)

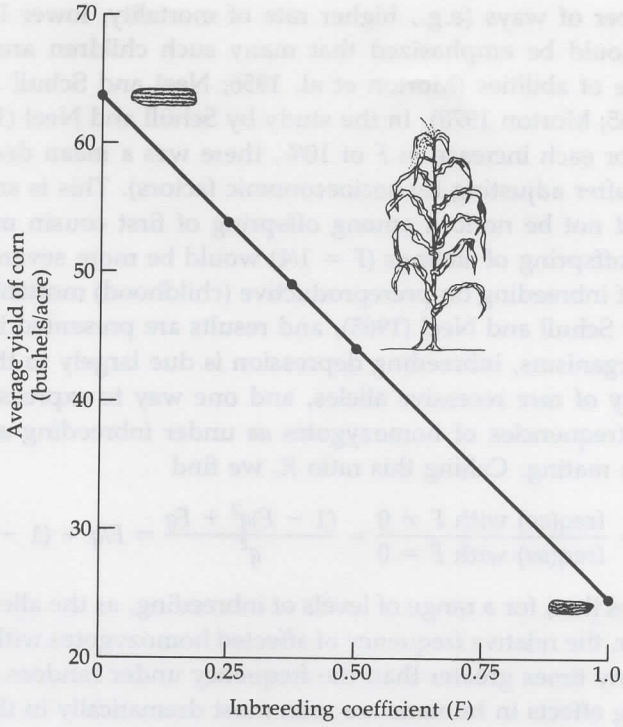


Figure 12. Decline of yield with inbreeding in corn (inbreeding depression). A linear inbreeding depression, as observed here, is expected to result from detrimental recessive alleles that are rendered homozygous by inbreeding. (Data from Neal 1935.)

diversity. In the case of allozymes, genetic diversity is due to common alleles that do not perceptibly impair viability or fertility when homozygous. In the case of inbreeding illustrated in Figure 11, we are dealing with rare alleles that are severely detrimental when homozygous. (That the alleles are rare is shown by the small proportion of lethal or near-lethal heterozygotes.) Figure 11 shows that natural populations contain considerable hidden genetic variation in the form of rare deleterious recessive alleles. (For more examples, see Dobzhansky et al. 1963; Crumpacker 1967; Mukai and Yamaguchi 1974.)

Detrimental effects of inbreeding (called **inbreeding depression**) are found in virtually all outcrossing species, and the more intense the inbreeding, the more harmful the effects (Figure 12). Inbreeding in humans is generally harmful, but the effect is difficult to measure because the degree of inbreeding is less than what is possible to achieve in experimental organisms, and the effects may vary from population to population. Nevertheless, children of first cousin matings are on the average less capable than noninbred children

in any number of ways (e.g., higher rate of mortality, lower I.Q. scores), though it should be emphasized that many such children are within the normal range of abilities (Morton et al. 1956; Neel and Schull 1962; Schull and Neel 1965; Morton 1978). In the study by Schull and Neel (1965), it was found that for each increase in F of 10%, there was a mean decrease in IQ of 6 points (after adjusting for socioeconomic factors). This is small enough that it would not be noticed among offspring of first cousin matings (F = 1/16), while offspring of siblings (F = 1/4) would be more severely affected. The effects of inbreeding on prereproductive (childhood) mortality were also examined by Schull and Neel (1965), and results are presented in Table 2.

In most organisms, inbreeding depression is due largely to the increased homozygosity of rare recessive alleles, and one way to express this is as a ratio of the frequencies of homozygotes *aa* under inbreeding as compared with random mating. Calling this ratio R, we find

$$R = \frac{\text{freq}(aa) \text{ with } F \neq 0}{\text{freq}(aa) \text{ with } F = 0} = \frac{(1 - F)q^2 + Fq}{q^2} = F/q + (1 - F) \qquad \textbf{(5.12)}$$

Table 3 shows that, for a range of levels of inbreeding, as the allele frequency q gets smaller, the relative frequency of affected homozygotes with inbreeding becomes many times greater than the frequency under random mating.

Inbreeding effects in humans are seen most dramatically in the enhanced frequency of genetic abnormalities due to harmful recessive alleles among the children of first cousin matings. The enhanced frequency of such conditions results from the genotype frequencies given in Table 1. If *a* denotes a rare deleterious recessive allele, then among the children of first cousin

Table 2. Effects of inbreeding on child mortality in Hiroshima 1948–1960.

	RELATIONSHIP OF PARENTS			
Time of child's death	Unrelated (F = 0)	Second cousins (F = 1/64)	First cousins once removed (F = 1/32)	First cousins (F = 1/16)
Stillborn	0.0198	0.0308	0.0140	0.0242
First month	0.0259	0.0385	0.0309	0.0347
Month 2–12	0.0183	0.0139	0.0112	0.0263
Year 1–12	0.0168	0.0256	0.0281	0.0273
Total	0.082	0.108	0.090	0.114
Sample size	1970	390	356	951

(From Schull and Neel 1965.)

Table 3. Effect of inbreeding on the incidence of homozygous recessives.

q	Sib mating (F = 1/4)	Uncle-niece (F = 1/8)	First cousins (F = 1/16)	Second cousins (F = 1/64)
0.100	3.25	2.12	1.56	1.14
0.050	5.75	3.38	2.19	1.30
0.010	25.75	13.38	7.19	2.55
0.005	50.75	25.88	13.44	4.11
0.001	250.75	125.88	63.44	16.61

$$\text{Ratio of } \frac{aa \text{ with inbreeding}}{aa \text{ under Hardy-Weinberg}} = \frac{q^2(1 - F) + qF}{q^2} = F/q + (1 - F)$$

matings the frequency of *aa* will be $q^2(1 - 1/16) + q(1/16)$, because for these children $F = 1/16$. If the proportion of first cousin matings in a population is c, then the total frequency of affected children will be $q^2(1 - c)$ (for the matings between nonrelatives) + $c[q^2(1 - 1/16) + q(1/16)]$ (for the first cousin matings). The proportion of all affected children that have first cousin parents will therefore equal the frequency of *aa* children from first cousin matings — $c[q^2(1 - 1/16) + q(1/16)]$ — divided by the total frequency of *aa* children given earlier. A little algebra shows that the proportion K of *aa* children with first cousin parents is

$$K = c(1 + 15q)/(c + 16q - cq) \qquad \textbf{(5.13)}$$

Curves of this equation are shown in Figure 13 for $c = 0.01$ (the approximate value of c for Caucasians in the United States) and $c = 0.06$ (the approximate value in Japan). Also graphed are data on several rare autosomal recessive conditions in Caucasians and Japanese. If there were no inbreeding effect, the proportion of affected children having **consanguineous** (related) parents would simply equal the proportion of consanguineous matings (0.01 in the case of Caucasians, 0.06 for Japanese). There is clearly a dramatic inbreeding effect, and the rarer the frequency of the deleterious recessive allele, the greater the effect.

REGULAR SYSTEMS OF INBREEDING

In plant and animal breeding, it is often useful to know how rapidly the inbreeding coefficient increases when a strain is propagated by a regular system of mating, such as repeated self-fertilization, sib mating, or backcross-

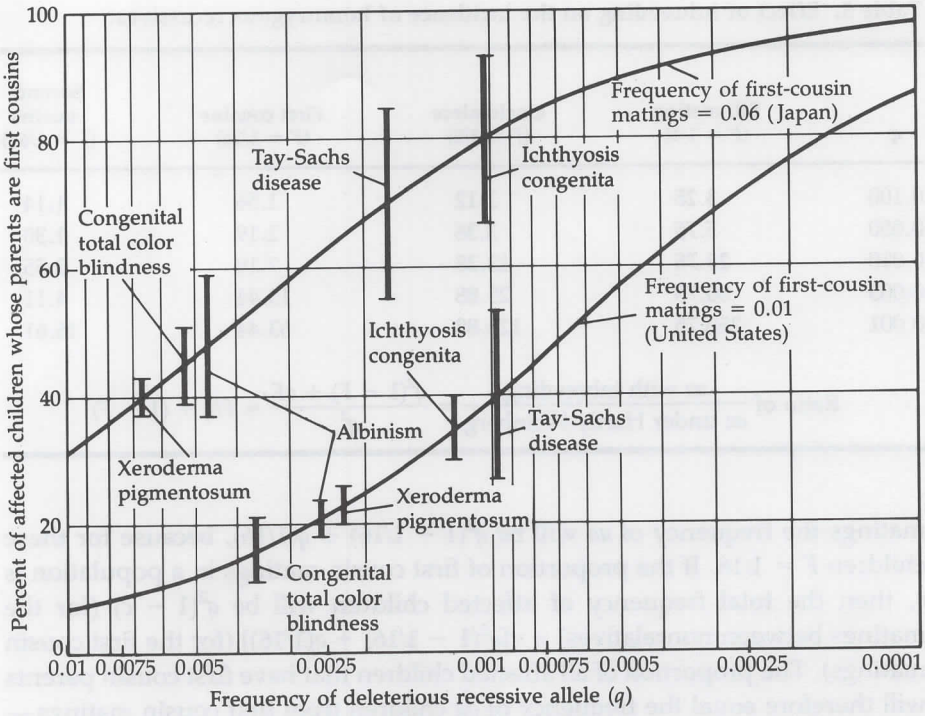


Figure 13. Curve of $c(1 + 15q)/(c + 16q - cq)$ for c (the proportion of first cousin matings) = 0.06 (Japan) and $c = 0.01$ (U.S.), against q (allele frequency of deleterious recessive). The ordinate is the proportion of affected children that have first cousin parents. Vertical bars represent 95% confidence intervals in the estimates. Note that this proportion increases with the rarity of the harmful allele. Data pertaining to various diseases due to rare autosomal recessives are from Morton (1961).

ing to a standard strain. The reasoning involved in calculating the inbreeding coefficient for any generation is illustrated in Figure 14 for repeated self-fertilization (for repeated sib mating, see Box D). In this figure, the labels $t - 1$ and t refer to the individuals after $t - 1$ and t generations of self-fertilization. The loop around the individual in generation $t - 1$ designates the probability that the two indicated alleles are identical by descent. Here the formula in Equation 5.4 applies with only one path and only one individual in that path, so

$$F_t = (1/2)^1(1 + F_{t-1}) \qquad \textbf{(5.14)}$$

where F_t is the inbreeding coefficient in generation t. This equation is easy to solve in terms of the quantity $1 - F_t$, which is often called the *panmictic index*. (Panmixia is a synonym for random mating.) Multiplying both sides of the equation for F_t by -1 and then adding $+1$ to each side leads to

$$
\begin{aligned}
1 - F_t &= 1 - (1/2)(1 + F_{t-1}) \\
&= 1 - (1/2) - (1/2)F_{t-1} \\
&= (1/2)(1 - F_{t-1})
\end{aligned}
\tag{5.15}
$$

or

$$1 - F_t = (1/2)^t(1 - F_0)$$

where F_0 is the inbreeding coefficient in the initial generation. Self-fertilization therefore leads to an extremely rapid increase in the inbreeding coefficient. When $F_0 = 0$, then $F_1 = 1/2$, $F_2 = 3/4$, $F_3 = 7/8$, $F_4 = 15/16$, and so on. Since $F_t = (H_0 - H_t)/H_0$, Equation 5.15 can be written in terms of heterozygosity as

$$H_t = (1/2)\, H_{t-1} \tag{5.16}$$

which shows directly that the heterozygosity halves each generation under recurrent selfing. How F increases when $F_0 = 0$ under self-fertilization and several other regular systems of mating is shown in Figure 15. If two lines are maintained by self-fertilization (or any other form of close inbreeding) for many generations, F will approach 1. However, if two different highly inbred lines are intercrossed, the coefficient of inbreeding of the offspring instantly returns to zero. Therefore, although many generations can be required to build up high values of the inbreeding coefficient, a single round of outcrossing wipes it out.

Many plants reproduce predominantly by self-fertilization, including crop plants such as soybeans, sorghum, barley, and wheat. As expected of highly self-fertilizing species, each individual plant is highly homozygous for alleles such as those determining allozymes. Yet the proportion of polymorphic genes is comparable to that found in outcrossing species; self-fertilization

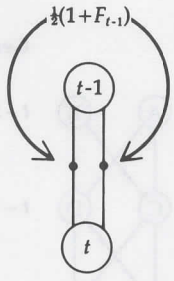


Figure 14. Increase of F due to continued self-fertilization. The individual in generation t is the offspring of self-fertilization of the individual in generation $t - 1$. The loop shows that $F_t = (1/2)(1 + F_{t-1})$.

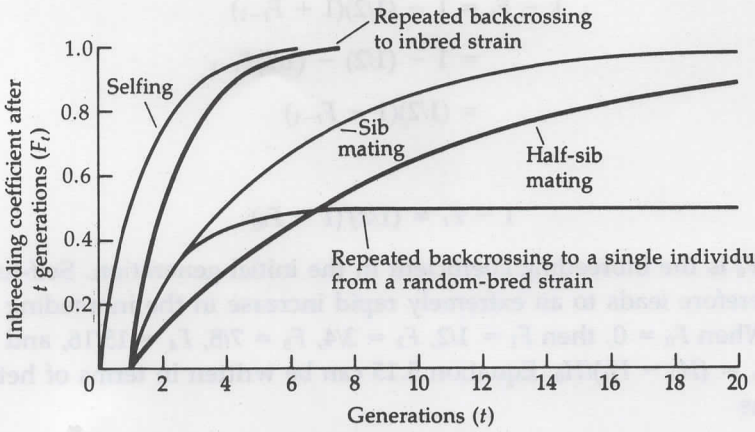


Figure 15. Theoretical increase in inbreeding coefficient F for regular systems of mating: selfing, sib mating, half-sib mating, and repeated backcrossing to a random-bred individual. The repeated backcrossing results in an equilibrium $\hat{F} = 1/2$, because at most 1/2 of the alleles from the population can be identical by descent to any single allele drawn from the random-bred backcrossing parent.

D Inbreeding Coefficient with Recurrent Sib Mating

The pedigree in Figure 16 shows three generations of recurrent brother–sister mating. With this mating scheme, an individual can inherit a pair of genes that are identical by descent if an individual in generation $t - 2$ was autozygous and passed both alleles to the same individual in generation t, or if two individuals in generation $t - 1$ passed the same allele to an individual in generation t. The easiest way to calculate the recurrence equation for F is to make use of the coefficient of coancestry. First note that

$$F_t = F_A = F_{BC}$$

The inbreeding coefficient of individual A is the probability that A has a pair of alleles

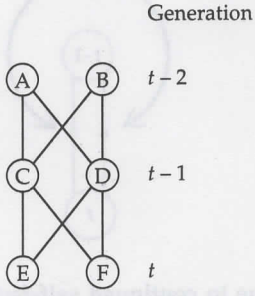


Figure 16. Pedigree for recurrent sib mating.

identical by descent, and this is equal to the probability that B and C transmit a pair of alleles identical by descent (which is the coefficient of coancestry F_{BC}). A coefficient of coancestry can be decomposed in the following way:

$$F_{BC} = F_{B(DE)}$$

This means that the probability that B and C transmit identical alleles is the same as the probability that an allele from B is identical to an allele transmitted by either of C's parents. This can be expanded as

$$F_{B(DE)} = (1/2)(F_{BD} + F_{BE})$$

Now we continue working backwards in time, expanding the right hand side of the above equation:

$$\begin{aligned} F_{B(DE)} &= (1/2)(F_{(DE)D} + F_{(DE)E}) \\ &= (1/2)[(1/2)(F_{DD} + F_{ED}) + (1/2)(F_{DE} + F_{EE})] \end{aligned}$$

and because $F_{DE} = F_{ED}$ and $F_{DD} = F_{EE}$, this is

$$F_t = (1/2)(F_{DD} + F_{DE})$$

We know that $F_{DE} = F_{t-1}$, by the rule that the coefficient of inbreeding in a given generation is the coefficient of coancestry of a random pair of individuals from the previous generation. The expression F_{DD} is the probability that D transmits a pair of alleles that are identical by descent. If D has alleles α_1 and α_2, then half the time the same allele is drawn twice (in which case $F = 1$), and half the time alleles α_1 and α_2 are drawn (and they are identical by descent with probability F_{t-2}). Putting these together,

$$F_{DD} = (1/2)(1 + F_{t-2})$$

Now substituting the expressions of F_{DD} and F_{DE} in terms of F_{t-1} and F_{t-2} into the expanded expression for F_t, we get

$$\begin{aligned} F_t &= (1/2)[(1/2)(1 + F_{t-2}) + F_{t-1}] \\ &= 1/4 + (1/2)F_{t-1} + (1/4)F_{t-2} \end{aligned}$$

With recurrent sib mating, F increases monotonically and asymptotically approaches $F = 1$. To see what happens to heterozygosity we can again make the substitution

$$F_t = (H_0 - H_t)/H_0$$

and obtain the recurrence equation in heterozygosity. This yields

$$H_t = (1/2)H_{t-1} + (1/4)H_{t-2}$$

An easy way to compare the rate of loss of heterozygosity under sib mating to the rate of loss under selfing is to calculate the proportion λ, by which H declines each generation. Then

$$\lambda = H_t/H_{t-1}$$

When we divide the equation for H_t by H_{t-1}, we get

$$H_t/H_{t-1} = \lambda = (1/2)(H_{t-1}/H_{t-1}) + (1/4)(H_{t-2}/H_{t-1})$$

$$= 1/2 + 1/(4\lambda)$$

The solution for λ is $\lambda = (1 + \sqrt{5})/4 = 0.80902$. This means that heterozygosity is lost at a rate of $1 - \lambda = 0.19098$, or about 19% per generation, with recurrent sib mating.

simply reorganizes the genetic variation into homozygous genotypes. On the other hand, self-fertilizing species contain fewer deleterious recessives than do outcrossing species, presumably because the increased homozygosity permits harmful recessives to be eliminated from the population by natural selection. One other important point about naturally self-fertilizing species: the high homozygosity of all genes implies that recombination rarely results in new gametic types not already present in the parent. Predominance of selfing in a species, therefore, has the effect of retarding the approach to linkage equilibrium, the reason being that the approach to linkage equilibrium is through recombination in double heterozygotes (*AB/ab* and *Ab/aB* in the case of two alleles of each gene), and with extreme inbreeding, such double heterozygotes are rare. Indeed, the most extreme examples of linkage disequilibrium have been found in predominantly self-fertilizing species such as barley (*Hordeum vulgare*) and wild oats (*Avena barbata*).

Clegg et al. (1972) provide an example of extreme linkage disequilibrium between two unlinked ($r = 1/2$) esterase genes in *Hordeum vulgare*, a species that regularly undergoes more than 99% self-fertilization. The population had originated as a complex cross and had been maintained for 26 generations under normal agricultural conditions without conscious selection. At the *esterase-B* locus, they distinguished two alleles, B_1 and B_2; at the *esterase-D* locus, they distinguished two alleles, D_1 and D_2. They observed the following numbers of each of the gametic types, although for practical purposes these numbers also refer to homozygous genotypes because there is such close inbreeding.

$B_1 D_1$	1501	(1642.6)
$B_1 D_2$	754	(613.7)
$B_2 D_1$	720	(577.1)
$B_2 D_2$	74	(215.6)

(The numbers in parentheses are the expected numbers based on the assumption of linkage equilibrium, calculated exactly as for the *MNSs* example given earlier.) The χ^2 value in this case is 172.7 with one degree of freedom. The associated probability is much less than 0.0001, so there certainly is significant linkage disequilibrium. For the above data, the linkage disequili-

brium parameter is $D = -0.046$, which is about 66% of its theoretical minimum. The consequences of inbreeding on a pair of genes are discussed further in the section on inbreeding and two loci (p. 263).

One of the dramatic successes of plant breeding has come from the crossing of inbred lines to produce high-yielding hybrid corn. Yield of a genetically heterogeneous outcrossing variety of corn can be improved by selecting the plants with the highest yields in each generation to be the progenitors of the next generation; however, such artificial selection results in only gradual improvement (see Chapter 8). If a large number of self-fertilized lines are established from a heterogeneous population, each line declines in yield as inbreeding proceeds, due to the forced homozygosity of deleterious recessives. Many lines become so inferior that they have to be discontinued. Self-fertilized lines are not likely to become homozygous for exactly the same set of deleterious recessives, however, so when different lines are crossed to produce a hybrid, the hybrid becomes heterozygous for these genes. Alleles favoring high yield in corn are generally dominant, and there may also be genes in which the heterozygous genotypes have a more favorable effect on yield than do the homozygous genotypes; in any case, the hybrid has a much higher yield than either inbred parent. The phenomenon of enhanced hybrid performance is called **hybrid vigor** or **heterosis**. In practice, inbred lines are crossed in many combinations to identify those that produce the best hybrids. Yields of hybrid corn are typically 15% to 35% greater than yields of outcrossing varieties, and the successful introduction of hybrid corn has been remarkable. Virtually all corn acreage in the United States today is planted with hybrids, as compared to 0.4% of the acreage in 1933 (Sprague 1978).

THE MIXED MATING MODEL (PARTIAL SELFING)

Many plants are capable of both outcrossing and self-fertilization, and they have evolved a diversity of mechanisms to regulate the fraction of seeds that are produced by the two mechanisms of fertilization. Because the rate of selfing varies widely among species, the population genetic consequences of mixed outcrossing and selfing are of great interest and importance. The mixed mating or partial selfing model is relatively easy to analyze because populations can be considered as a mixture of selfed and outcrossed individuals. In the discussion of genotypic frequencies with inbreeding, we noted that the population could always be considered as a panmictic group, with frequency $(1 - F)$, and as a totally inbred group, with frequency F.

Let S be the fraction of the population that results from self-fertilization, and $1 - S$ the fraction from random mating. The parameter S is usually called the selfing rate, and sometimes the parameter $T = 1 - S$ is used to measure

the outcrossing rate. With repeated self-fertilization, we saw that the inbreeding coefficient changes according to

$$F_{t+1} = (1/2)(1 + F_t) \tag{5.17}$$

Since the fraction S of the population undergoes self-fertilization, and the fraction $(1 - S)$ remains randomly mating ($F = 0$), the inbreeding coefficient when both occur is

$$F_{t+1} = S[(1/2)(1 + F_t)] + (1 - S)(0) \tag{5.18a}$$

or

$$F_{t+1} = (S/2)(1 + F_t) \tag{5.18b}$$

Therefore, the inbreeding coefficient increases at a rate equal to S *times the rate for complete selfing.* F does not increase to $F = 1$, and we can solve for the equilibrium value of F by setting $F_{t+1} = F_t$:

$$F_{t+1} = F_t = (S/2)(1 + F_t) \tag{5.19}$$

After some easy algebraic rearranging, this yields

$$\hat{F} = S/(2 - S) \tag{5.20}$$

The mixed mating model results in an intermediate level of inbreeding determined by the fraction of the population that self pollinates. This result

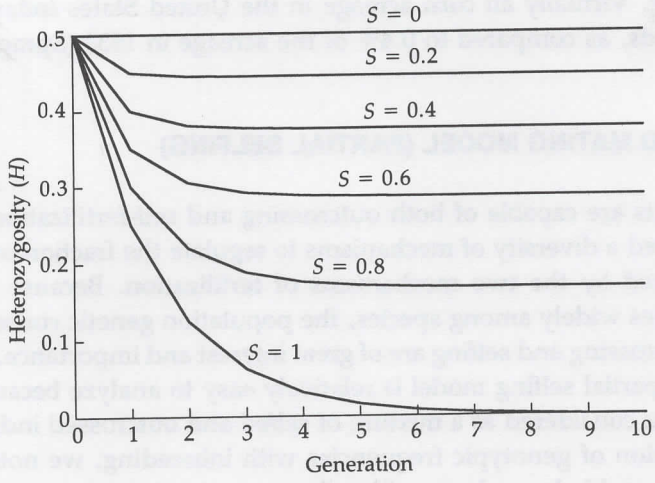


Figure 17. Heterozygosity in the mixed mating model, where a fraction S of the population self-fertilizes and the remainder mate at random. $S = 0$ in an ordinary panmictic population, and remains in Hardy-Weinberg proportions, while $S = 1$ results in a halving in heterozygosity each generation. In all cases the allele frequency does not change, and the equilibrium is $\hat{F} = S/(2 - S)$.

assumes that the plants that self-fertilize are a random sample, so that it is no more likely that a selfed plant produces selfing progeny than an outcrossed plant. Equilibrium genotype frequencies can be calculated using Equations 5.8 and substituting $S/(2 - S)$ for the value of F. Using the relation $F_t = (H_0 - H_t)/H_0$, Equation 5.19 can be written as

$$H_{t+1} = H_0(1 - S) + (S/2)H_t \qquad \textbf{(5.21)}$$

The decline in heterozygosity in populations with partial selfing is plotted in Figure 17 for several rates of selfing. Box E shows the calculation of genotype frequencies in a natural plant population that undergoes partial selfing.

One-Locus Genotype Frequencies Under the Mixed Mating Model

E

For a population of plants that are self-compatible (i.e., able to undergo self-fertilization), if each plant undergoes self-fertilization a fraction S of the time and otherwise mates randomly, then F very quickly attains the value $F = S/(2 - S)$. *Phlox cuspidata* is self-compatible, and for this species the amount of self-fertilization is estimated at approximately $\langle S \rangle = 0.78$ (Levin 1978); thus we can estimate the inbreeding coefficient as $\langle F \rangle = 0.78/(2 - 0.78) = 0.64$. In a Texas population of *P. cuspidata*, Levin (1978) found two alleles of the gene for phosphoglucomutase-2, designated $Pgm\text{-}2^a$ and $Pgm\text{-}2^b$. In a sample of 35 individuals, he found 15 $Pgm\text{-}2^a/Pgm\text{-}2^a$, 6 $Pgm\text{-}2^a/Pgm\text{-}2^b$, and 14 $Pgm\text{-}2^b/Pgm\text{-}2^b$ individuals. Are these numbers consistent with the estimate $F = 0.64$? (Note: The χ^2 here has one degree of freedom because only the allele frequency is estimated from the data; if F also were estimated from the data, rather than being calculated independently from the degree of self-fertilization, then there would be zero degrees of freedom and a goodness-of-fit test would not be possible.)

ANSWER Allele frequencies of $Pgm\text{-}2^a$ and $Pgm\text{-}2^b$ are estimated as $(30 + 6)/70 = 0.514$ and $1 - 0.514 = 0.486$, respectively. The hypothesis is that $F = 0.64$, thus $1 - F = 0.36$. Expected numbers of the three genotypes, given in the order above, are $[(0.514)2(0.36) + (0.514)(0.64)](35) = 14.8$; $[2(0.514)(0.486)(0.36)](35) = 6.3$; and $[(0.486)2(0.36) + (0.486)(0.64)](35) = 13.9$.

The χ^2 is 0.02 with one degree of freedom, and the associated probability is about 0.96. The fit to the inbreeding model is excellent.

INBREEDING AND TWO LOCI

As mentioned above, inbreeding affects the array of gametic frequencies in a population, even though it does not by itself affect allele frequencies. To see the effect of inbreeding on a pair of genes, consider a population started

with genotype *AaBb* only, where genes *A* and *B* are unlinked. The genotype *AaBb* produces the four gametes *AB*, *Ab*, *aB*, and *ab* in equal frequency, so after one generation of selfing, (1) each of the four doubly homozygous genotypes has frequency (1/4)(1/4) = 1/16, (2) each of the four single heterozygotes has frequency 2(1/4)(1/4) = 1/8, because they can get their two different gametes in either of two orders (*Ab* and *AB* for instance), and (3) the double heterozygotes have frequency 1/4, because they can get either gametes *AB* and *ab*, or *Ab* and *aB*, again in either of two orders. Over successive generations, the frequency of double heterozygotes decreases by a factor of 1/4 each generation, until they vanish. Among selfed offspring of single heterozygotes, there are 1/4 of each of the two homozygous classes and 1/2 heterozygotes, so the single heterozygotes decrease in frequency also. Eventually, there remain just four genotypes, each a double homozygote with a frequency equal to the initial gamete frequency. The process of selfing with two genes is illustrated in Figure 18.

Calculation of the ten two-locus genotype frequencies in an inbreeding population can be very complex, and a nomenclature for handling the problem has been devised by Weir and Cockerham (1973). The reason two-locus inbreeding becomes so complex is that genes can be identical by descent in many different ways. For example, two copies of allele α_1 in an individual might derive from one common ancestor and alleles β_1 and β_2 at the other locus might not be identical by descent; or the individual could have two α_1 alleles and two β_1 alleles that are identical by descent, either through the same or different common ancestors. These and other possibilities are summarized by **digenic measures of descent**, which represent the probabilities of all possible forms of identity. The essence of this theory can be made clear by comparing one such measure with a statistic that relates it to the normal inbreeding coefficient.

First recall that the one-locus inbreeding coefficient *F* represents the probability that a pair of uniting gametes is identical by descent, and all formulae for the change in *F* are independent of linkage. The two-locus measure that is most useful is denoted F_{11}, and it is the probability that an individual receives a pair of alleles that are identical by descent at *both* loci. This event must be less likely than receiving a pair of alleles identical by descent at one locus, so $F_{11} \leq F$. A useful parameter that represents the relation between the one- and two-locus inbreeding coefficients is the **identity disequilibrium**, which is defined as

$$\eta_{11} = F_{11} - F^2 \qquad \textbf{(5.22)}$$

The identity disequilibrium η_{11} is a disequilibrium measure in the sense that it indicates the departure of two-locus genotype frequencies from the product of one-locus genotype frequencies, much in the way the linkage disequili-

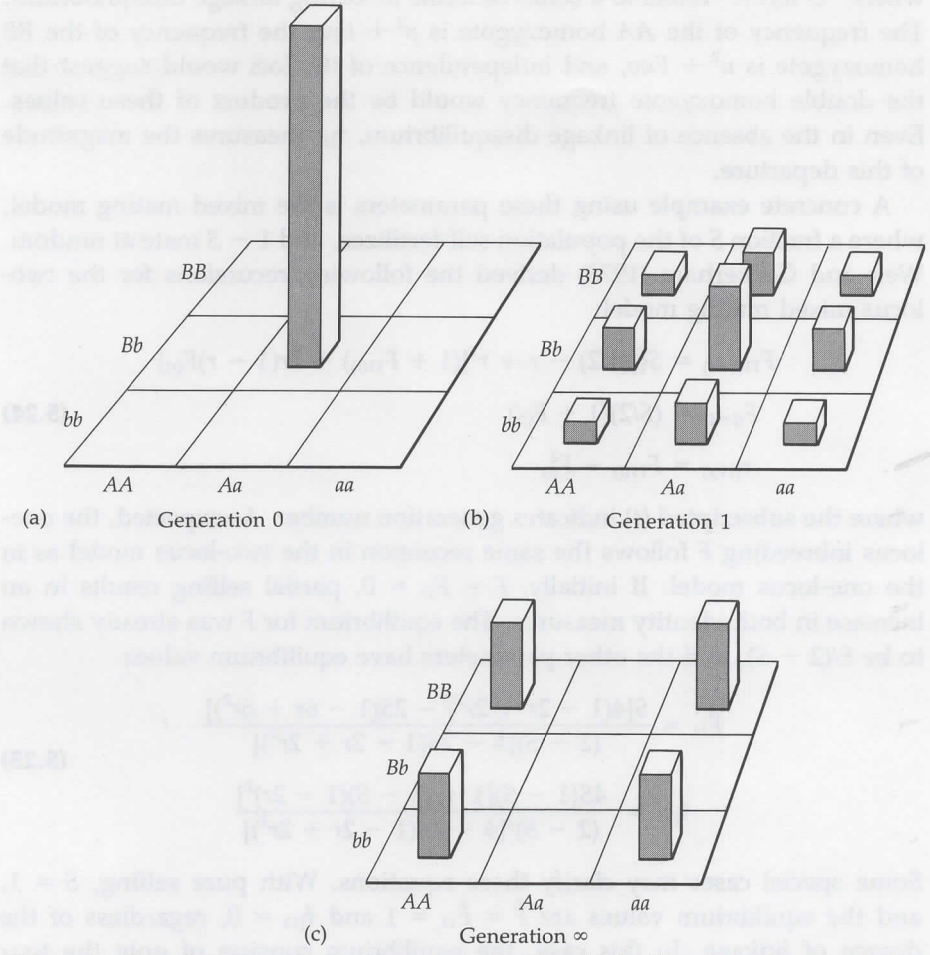


Figure 18. Two-locus genotype frequencies with self-fertilization in a population started with a double heterozygote. Genes *A* and *B* are unlinked in this example.

brium parameter measures departures of gametic frequencies from the product of allele frequencies. As we will show, even if the genes are unlinked, a statistical interdependence of identity dictates that the probability of both genes having alleles that are identical by descent is greater than the product of the probabilities of identity of each gene separately. This means that $F_{11} \geq F^2$, and hence $\eta_{11} \geq 0$.

To relate the descent measures to genotype frequencies, let the frequency of alleles A, a, B, and b be p, q, u, and v respectively. If P_{AB}^{AB} is the frequency of the AB/AB genotype, then

$$P_{AB}^{AB} = (p^2 + Fpq)(u^2 + Fuv) + \eta_{11}pquv + D \text{ terms} \qquad \textbf{(5.23)}$$

where "D terms" refers to a series of terms involving linkage disequilibrium. The frequency of the AA homozygote is $p^2 + Fpq$, the frequency of the BB homozygote is $u^2 + Fuv$, and independence of the loci would suggest that the double homozygote frequency would be the product of these values. Even in the absence of linkage disequilibrium, η_{11} measures the magnitude of this departure.

A concrete example using these parameters is the mixed mating model, where a fraction S of the population self-fertilizes, and $1 - S$ mate at random. Weir and Cockerham (1973) derived the following recursions for the two-locus mixed mating model:

$$\begin{aligned} F_{11(t+1)} &= S\{[(1/2) - r + r^2](1 + F_{11(t)}) + 2r(1 - r)F_{(t)}\} \\ F_{(t+1)} &= (S/2)(1 + F_{(t)}) \\ \eta_{11(t)} &= F_{11(t)} - F^2_{(t)} \end{aligned} \tag{5.24}$$

where the subscripted (t) indicates generation number. As expected, the one-locus inbreeding F follows the same recursion in the two-locus model as in the one-locus model. If initially, $F = F_{11} = 0$, partial selfing results in an increase in both identity measures. The equilibrium for F was already shown to be $S/(2 - S)$, and the other parameters have equilibrium values

$$\begin{aligned} \hat{F}_{11} &= \frac{S[4(1 - 2r + 2r^2) - 2S(1 - 6r + 6r^2)]}{(2 - S)[4 - 2S(1 - 2r + 2r^2)]} \\ \hat{\eta}_{11} &= \frac{4S(1 - S)[1 + (1 - S)(1 - 2r)^2]}{(2 - S)^2[4 - 2S(1 - 2r + 2r^2)]} \end{aligned} \tag{5.25}$$

Some special cases may clarify these equations. With pure selfing, $S = 1$, and the equilibrium values are $\hat{F} = \hat{F}_{11} = 1$ and $\hat{\eta}_{11} = 0$, regardless of the degree of linkage. In this case, the equilibrium consists of only the four double homozygotes, and their frequencies are the products of the respective allele frequencies. With no selfing, $S = 0$, and $\hat{F} = \hat{F}_{11} = \hat{\eta}_{11} = 0$, and the population contains all genotypes with frequencies equal to the products of the respective allele frequencies.

In the case of absolute linkage ($r = 0$), the equilibrium given by Equations 5.25 are $\hat{F} = \hat{F}_{11} = S/(2 - S)$, and $\hat{\eta}_{11} = [S/(2 - S)][1 - S/(2 - S)]$. Because recombination breaks down associations of alleles, it is not surprising that the case of $r = 0$ yields the greatest η_{11}, reflecting the greatest possible excess of double homozygote frequency over the product of single locus homozygote frequencies. Finally, with $r = 1/2$, the equilibrium values are

$$\begin{aligned} \hat{F} &= S/(2 - S) \\ \hat{F}_{11} &= S(2 + S)/[(2 - S)(4 - S)] \\ \hat{\eta}_{11} &= 4S(1 - S)/[(2 - S)^2(4 - S)] \end{aligned} \tag{5.26}$$

Here is a big surprise of the two-locus mixed mating model: even if the two loci are unlinked, the identity disequilibrium is greater than zero. This result is sometimes called the Bennett-Binet effect, and it implies a tendency for homozygosity to be concentrated in some individuals to a degree greater than expected from one-locus genotype frequencies (Bennett and Binet 1956). The phenomenon is related to an observation in the infinite alleles model for a pair of loci, which generates correlations in heterozygosity even between unlinked loci (Chapter 3). Figure 19 shows the identity disequilibrium for the range of S with several different levels of linkage. Note that the greatest identity disequilibrium is obtained for $r = 0$, but that even with $r = 1/2$, η_{11} can exceed 0.15. When $S = 0$ or $S = 1$, the identity disequilibrium is zero.

To this point we have discussed only departures of genotype frequencies from panmictic proportions, but partial selfing has an effect on linkage disequilibrium as well. Although D does eventually go to zero under this model, inbreeding retards the rate of decay of linkage disequilibrium. Weir et al. (1972) determined the recurrence equation for linkage disequilibrium in the two-locus mixed mating model:

$$D' = (1/2)\{1 - r + (1/2)S + \sqrt{[1 - r + (1/2)S]^2 - 2S(1 - 2r)}\}(D) \quad \textbf{(5.27)}$$

Thus the linkage disequilibrium D decreases according to a geometric series. With no selfing, Equation 5.27 gives the familiar $D' = (1 - r)D$; and when

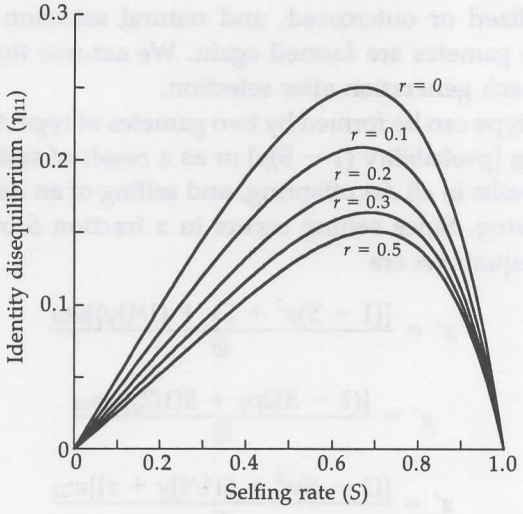


Figure 19. The identity disequilibrium η_{11} in the two-locus mixed mating model, plotted over the range of selfing rate S for several values of recombination frequency r. η_{11} measures the degree of excess of double homozygote frequency over the product of the one-locus homozygote genotype frequencies, and hence measures the clustering of homozygosity (and heterozygosity) in individuals.

$r = 1/2$, Equation 5.27 implies that $D' = (1/2)(1 + S)D$. Increasing recombination increases the rate at which linkage disequilibrium disappears, while increasing the rate of selfing retards the decrease in D. With free recombination in a totally self-fertilizing population, $D' = D$ and the linkage disequilibrium does not decrease. This result is expected because, at equilibrium with complete selfing, the four double homozygotes each have frequencies equal to the initial gametic frequencies, and since the equilibrium gametic frequencies stay the same, D stays the same.

SELECTION AND INBREEDING

Although inbreeding by itself does not change allele frequencies, it can have a profound effect in combination with natural selection because of the increase in frequency of homozygous genotypes. Consider a population of diploids with genotypes AA, Aa, and aa in respective frequencies x, y, and z, with allele frequencies of A and a equal to p and q. Let the viabilities of genotypes AA, Aa, and aa be w_{11}, w_{12}, and w_{22}, respectively, and suppose that a fraction S of the population self-fertilizes and the fraction $1 - S$ mate at random (the mixed mating model). Since the genotypes do not occur in Hardy-Weinberg proportions, the recurrence equations cannot be written in terms of the allele frequencies, but rather require the genotype frequencies. Each generation, gametes are formed and united according to whether the parent self-fertilized or outcrossed, and natural selection operates on the offspring before gametes are formed again. We assume that the population is censused in each generation after selection.

The AA genotype can be formed by two gametes of type A coming together from outcrossing [probability $(1 - S)p$] or as a result of selfing. Selfing of an AA genotype results in all AA offspring, and selfing of an Aa genotype results in 1/4 AA offspring. Since selfing occurs in a fraction S of the population, the recurrence equations are

$$x' = \frac{[(1 - S)p^2 + Sx + (1/4)y)]w_{11}}{\overline{w}}$$

$$y' = \frac{[(1 - S)2pq + S(1/2)y]w_{12}}{\overline{w}} \qquad \textbf{(5.28)}$$

$$z' = \frac{[(1 - S)q^2 + S(1/4)y + z)]w_{22}}{\overline{w}}$$

where $\overline{w}$ is the sum of the numerators on the right side, and can be written

$$\overline{w} = (1 - S)\overline{w}_0 + S\{w_{11}[x + (1/4)y] + w_{12}(1/2)y + w_{22}[(1/4)y + z)]\} \qquad \textbf{(5.29)}$$

where $\overline{w}_0$ is the mean fitness without selfing. As promised, the recurrence system for the mixed mating model with viability selection is complex. (For a detailed analysis, see Workman and Jain 1966; Weir 1970; Kimura and Ohta 1971.)

We will summarize the qualitative behavior of the model, referring to Figures 20, 21, and 22. In Figure 20, plots of Δp vs. p are presented for a range of selfing rates. The classic viability model is $S = 0$. As S increases, the equilibrium moves towards the more fit of the two homozygotes. This is intuitive because, as the selfing rate increases, the frequency of both homozygous classes increases. In the extreme case $S = 1$, the population contains only the two homozygous classes, and if natural selection favors one of them, the respective allele will increase to fixation. Taking these observations together, it is not surprising that there is an intermediate value of S at which the population evolves to fixation even though heterozygotes are favored.

Figure 21 shows the region of the parameter space that admits a stable polymorphism. Note that as S increases, the region contracts closer to the line defined by $w_{11} = w_{22}$. This implies that polymorphism can be maintained with high rates of selfing provided the selection pattern is nearly symmetric, but in general, inbreeding makes it less likely for viability selection to maintain a polymorphism.

Figure 22 shows the mean fitness in the mixed mating model. When selfing occurs, not only does the equilibrium shift away from the maximum in mean

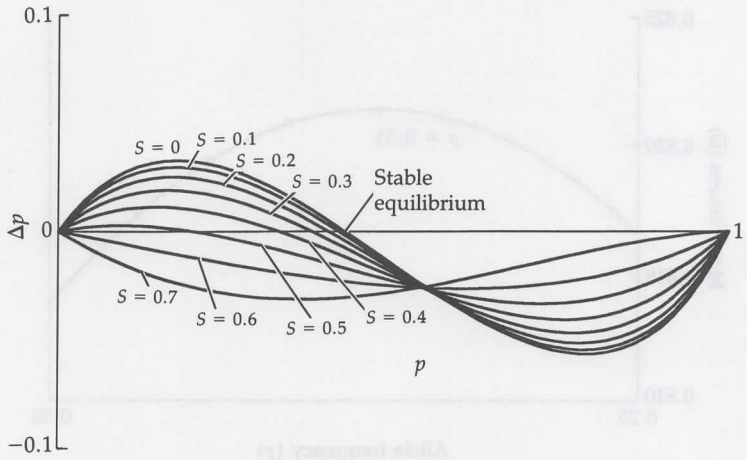


Figure 20. Δp versus p for the mixed mating model with viability selection. S is the fraction of the population that self-fertilizes. As the selfing rate increases, the equilibrium allele frequency moves toward, and finally joins, the fixation $p = 0$, at which point there is no internal equilibrium.

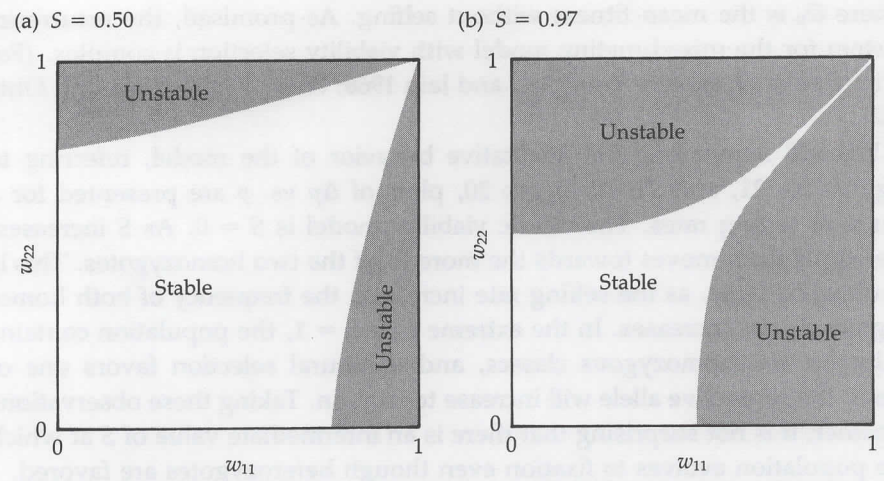


Figure 21. Parameter space admitting a stable equilibrium under the mixed mating model with viability selection. In these cases, $w_{12} = 1$, so the entire square area represents overdominance, and in the absence of selfing, a polymorphism would be guaranteed. Note that the unshaded region, representing a stable polymorphism, becomes narrower as S (the selfing rate) increases, implying that there must be a greater degree of overdominance, and greater symmetry to the selection.

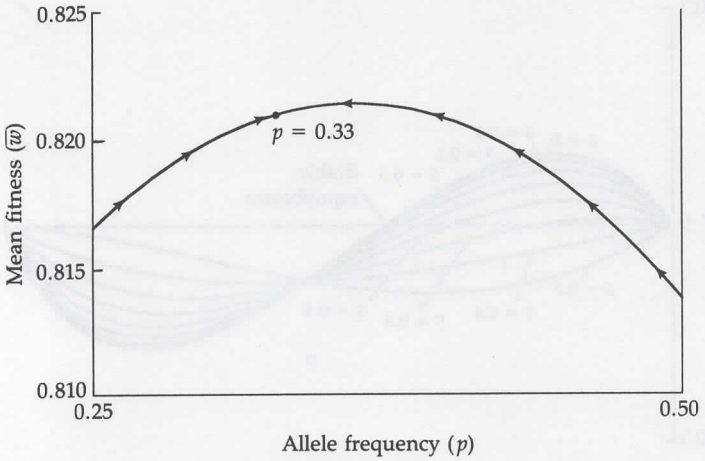


Figure 22. Under the mixed mating model, $\overline{w}$ is not maximized. In this example, $w_{11} = 0.6$, $w_{12} = 1$, and $w_{22} = 0.7$, with a selfing rate of $S = 0.4$. The equilibrium in this case is $\hat{p} = 0.33$.

fitness, but the entire curve for the mean fitness is different because the relative frequencies of genotypes change with inbreeding.

A special case of the mixed mating model occurs when one locus is subject to selection and a linked neutral locus is also segregating in the population (Cockerham and Rawlings 1967; Strobeck 1979). In this case, overdominance at the selected locus generates an excess of heterozygotes not only at the selected locus but at all neutral loci in the genome, regardless of linkage to the selected locus. This result emphasizes again that inbreeding affects the genotype frequencies of all genes, regardless of linkage. Because inbreeding generates a sort of correlation of heterozygosities across genes, if one gene is subject to overdominant selection, all genes will appear to have overdominant selection operating. This phenomenon makes it difficult to attribute selective effects to individual genes in partially selfing plant populations.

ASSORTATIVE MATING

Although random mating is the most important mating system in many natural populations, certain departures from random mating are also important. When choice of mates is based on phenotype, mating is **assortative**. In positive assortative mating (often called simply assortative mating), individuals tend to choose mates that are phenotypically like themselves. In negative assortative mating (also called disassortative mating), individuals tend to choose mates that are phenotypically unlike themselves. Of course, even with random mating, some mating pairs are phenotypically similar or dissimilar, so assortative mating refers only to those situations in which mating partners are phenotypically more similar or dissimilar than would be expected by chance in a random mating population.

Positive Assortative Mating

At first sight, positive assortative mating bears a strong resemblance to inbreeding. When a trait is determined by one or a few genes, individuals that have similar phenotypes will also be similar genetically; and if genetically similar individuals are more likely to mate than would occur with random mating, then this is a form of inbreeding. As with inbreeding, one consequence of positive assortative mating is to increase the homozygosity of genes affecting the traits on which mate selection is based. By increasing the frequency of different multiple-locus homozygotes, the phenotypic variance in the population is also increased. Another similarity between assortative mating and inbreeding is that, while neither changes allele frequencies by itself,

together they can have a profound effect in the presence of selection.

However, there are several important differences between positive assortative mating and inbreeding. In the first place, inbreeding affects all genes, while assortative mating affects only those genes that are associated with the phenotypic traits affecting mating frequencies. This results in assortative mating having a smaller effect than inbreeding on the homozygosity of "background" genes. Similarly, because assortative mating keys on a few genes, those genes can be more strongly affected than with inbreeding. In particular, the variance among traits associated with assortative mating can be much greater than would occur with inbreeding.

The pollination biology of flowering plants provides an example of positive assortative mating. For example, when the length of time during which any individual plant flowers is short relative to the total duration of the flowering season, plants that flower early in the season are preferentially pollinated by other early flowering plants, and those that flower late are preferentially pollinated by other late flowering ones. Thus, there is positive assortative mating for flowering time.

In humans, positive assortative mating occurs for height, IQ score, and certain other traits, although the amount of assortative mating varies in different populations and does not occur at all in some. An important consequence of assortative mating for IQ is that it tends to increase the genetic variance for this trait. We will see in Chapter 8 that this results in an inflated estimate of heritability. As might be expected, positive assortative mating occurs for certain socioeconomic variables. This effect is seen in one study in the United States, in which the highest correlation found between married couples was in the number of rooms in their parents' homes.

The consequences of positive assortative mating are complex and depend on the number of genes that influence the trait in question, on the number of different possible alleles of the genes, on the number of different phenotypes, on the sex performing the mate selection, and on the criteria for mate selection. Rarely are traits with assortative mating determined by the alleles of a single gene. Most traits affected by assortative mating are polygenic, so reasonably realistic models of assortative mating tend to be rather complex. Nevertheless, single gene models are instructive in showing the tendency for assortative mating to reduce heterozygosity. The simplest case of complete positive assortative mating concerns one gene with two alleles and codominance. In this case, there are three genotypes AA, Aa, and aa, and each has a distinct phenotype. The only matings are $AA \times AA$, $Aa \times Aa$, and $aa \times aa$. The model is identical to the model of selfing, so the recurrence equations for the change in F and H apply. However, an important difference between the one-locus codominant model of assortative mating and selfing is that, as soon as any other genes are considered, the two results are very different.

For example, a second gene B that is neither linked to A nor associated with assortative mating will not have its heterozygosity reduced by half in each generation. Inbreeding, on the other hand, does reduce the heterozygosity of all loci by the same proportion.

Consider a trait determined by one gene with complete dominance, so that genotypes AA and Aa have phenotype X and genotype aa has phenotype Y. With complete positive assortative mating, the only two mating types will be X × X and Y × Y. Table 4 shows the four genotypic possibilities for mating types and their frequencies. In Table 4, the frequencies of AA, Aa, and aa are x, y, and z, respectively, where $x + y + z = 1$. The recurrence equation for each genotype is obtained by multiplying the frequency of each mating type by the proportion of offspring having the genotype, and summing the columns. The results are

$$x' = \frac{[x + (1/2)y]^2}{1 - z}$$

$$y' = \frac{y[x + (1/2)y]}{1 - z} \qquad \textbf{(5.30)}$$

$$z' = \frac{(1/4)z^2}{1 - z} + z$$

Since the first term in the equation for z' is positive, it is immediately clear that the frequency of genotype aa increases over time. Also, since the term $x + (1/2)y = p$ is an allele frequency, and $0 < p < 1$, the frequency of heterozygotes must decrease over time. Eventually the population consists only of the two homozygous genotypes with relative frequencies p and q. You should

Table 4. Complete positive assortative mating for a dominant phenotype.

		PROGENY		
Mating	**Frequency**	***AA***	***Aa***	***aa***
$AA \times AA$	$\frac{x^2}{1 - z}$	1	0	0
$AA \times Aa$	$\frac{2xy}{1 - z}$	1/2	1/2	0
$Aa \times Aa$	$\frac{y^2}{1 - z}$	1/4	1/2	1/4
$aa \times aa$	z	0	0	1
Total	1	$\frac{(x + 1/2y)^2}{1 - z}$	$\frac{y(x + 1/2y)}{1 - z}$	$\frac{(1/4)y^2}{1 - z} + z$

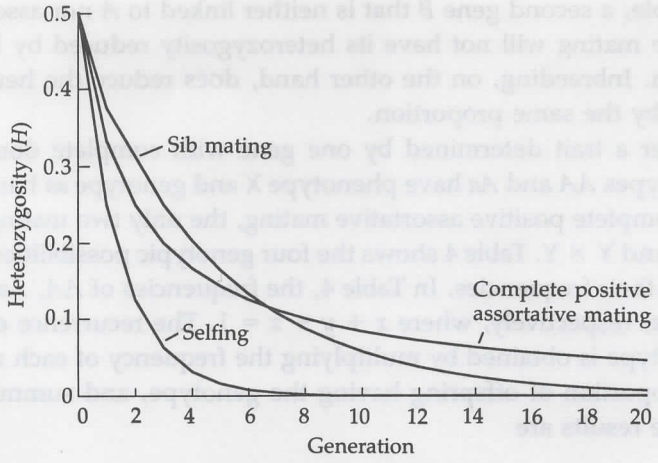


Figure 23. The decline in heterozygosity with positive assortative mating is initially more rapid than under sib mating but less than selfing. With positive assortative mating heterozygosity is lost more slowly in each succeeding generation.

use Equations 5.30 to verify that the allele frequencies do not change over time. With complete positive assortative mating, the heterozygosity $H = y$, and Equations 5.30 imply

$$H' = \frac{y[x + (1/2)y]}{x + (1/2)y + (1/2)y} = \frac{Hp}{p + (1/2)H} \qquad \textbf{(5.31)}$$

For the particular case $H_0 = 0.5$ and $p = 0.5$, heterozygosity is reduced by 1/3 in one generation, which is a greater reduction than with sib mating but less than with selfing. With assortative mating, in contrast to sib mating or selfing, the fraction by which H decreases each generation itself decreases (Figure 23). As H approaches zero, the term $p/[p + (1/2)H]$ in Equation 5.31 approaches 1, which means $H' \simeq H$. Contrast this result with selfing, in which the heterozygosity halves every generation even when H is small, and with sib mating, where the decline in heterozygosity is a constant proportion ($\lambda = 0.8090$) of the heterozygosity remaining each generation. Assortative mating decreases H at a rate slower than geometric because of the changes in frequencies of mating types.

Negative Assortative Mating

Negative assortative mating occurs when like phenotypes mate less frequently than expected with random mating. Negative assortative mating is particularly important in plants, where a wide array of assorting mechanisms has evolved. One mechanism is self-incompatibility, first described in tobacco by East and Mangelsdorf (1925). There are actually two distinct forms

of self-incompatibility: (1) gametophytic, in which pollen with an allele in common with the maternal plant fails to fertilize, and (2) sporophytic, in which the male and female diploid plants must be of different genotypes for successful fertilization. The simplest form of gametophytic self-incompatibility is illustrated with three alleles, S_1, S_2, and S_3. Since the incompatibility mechanism precludes homozygotes, only the three genotypes S_1S_2, S_1S_3, and S_2S_3 occur. Table 5 summarizes the mating types and the frequencies of the resulting offspring. The recurrence equations for this system are

$$\begin{aligned} x' &= (1/2)(1 - x) \\ y' &= (1/2)(1 - y) \\ z' &= (1/2)(1 - z) \end{aligned} \tag{5.32}$$

At equilibrium, $x' = x$, $y' = y$, $z' = z$, and therefore $\hat{x} = \hat{y} = \hat{z} = 1/3$, which is a stable equilibrium. The model is easily generalized to many self-incompatibility alleles, which is the usual situation in nature. With n alleles, the equilibrium frequencies (assuming no mutation, migration, random drift, or selection) are $1/n$ for each allele. In *Brassica*, self-incompatibility alleles have very different DNA sequences (Nasrallah et al. 1985); we will return to this phenomenon in Chapter 7.

Gametophytic incompatibility can be divided into two categories based on the reproductive success of females. The zygote elimination category occurs when the mating frequency between phenotypes P_i and P_j is $P_iP_j/\Sigma P_iP_j$, with the sum taken over all admissible matings. Karlin (1969) uses the term *mass action* for this model, since zygotes are not really eliminated from the population, and the mating frequencies are those expected with a model of

Table 5. Gametophytic self-incompatibility.

		PROGENY		
Mating type	Frequency	S_1S_2 (x)	S_1S_3 (y)	S_2S_3 (z)
$S_1S_2 \times S_1S_3$	x	0	1/2	1/2
$S_1S_2 \times S_2S_3$		0	1/2	1/2
$S_1S_3 \times S_1S_2$	y	1/2	0	1/2
$S_1S_3 \times S_2S_3$		1/2	0	1/2
$S_2S_3 \times S_1S_2$	z	1/2	1/2	0
$S_2S_3 \times S_1S_3$		1/2	1/2	0
Total	1	$(1/2)(y + z)$	$(1/2)(x + z)$	$(1/2)(x + y)$

mass action in which the genotypes are like particles bouncing around in space, mating when they collide. An example of negative assortative mating of the mass action type is *heterostyly*, which occurs in most species of primroses (*Primula*) and their relatives. In most primrose populations, there are approximately equal proportions of two types of flowers, one known as pin, which has a tall style and short stamens, and the other known as thrum, which has a short style and tall stamens. (In botanical terminology, the style carries the stigma, which is the organ that receives pollen, whereas the stamens carry the anthers, which produce pollen.) In heterostyly, insect pollinators that work high on the flowers pick up mostly thrum pollen and deposit it on pin stigmas, whereas pollinators that work low in the flowers pick up mostly pin pollen and deposit it on thrum stigmas. Negative assortative mating therefore occurs because pins mate preferentially with thrums.

Table 6 shows the mating table for the mass action model when only unlike genotypes mate. The recurrence equations are again obtained by summing the column of progeny frequencies weighted by the mating type frequency. The result is

$$\begin{aligned} x' &= xy/T \\ y' &= (xy + 2xz + yz)/T \\ z' &= yz/T \end{aligned} \tag{5.33}$$

where $T = 2(xy + yz + xz)$. Dividing the top equation by the bottom one shows that the ratio $x'/z' = x/z = K$ does not change over time. This feature implies that changes in genotype frequencies follow straight lines in De Finetti diagrams (Figure 24). To find the equilibrium, set $z' = z$, which implies $\hat{y} = T$. Solving for z and using the relations $x/z = K$ and $x + y + z = 1$, we obtain (Falk and Li 1969)

Table 6. Complete negative genotypic assortment (zygote elimination).

		PROGENY		
Mating	**Frequency**	*AA*	*Aa*	*aa*
$AA \times Aa$	xy	1/2	1/2	0
$AA \times aa$	xz	0	1	0
$Aa \times AA$	xy	1/2	1/2	0
$Aa \times aa$	yz	0	1/2	1/2
$aa \times AA$	xz	0	1	0
$aa \times Aa$	yz	0	1/2	1/2
Total	$2(xy + xz + yz)$	xy	$xy + 2xz + yz$	yz

$$\hat{z} = \frac{3(K + 1) - \sqrt{[(K + 1)^2 + 8K]}}{4(K^2 + K + 1)} \tag{5.34}$$

when $K = x/z = x_0/z_0$. Figure 24 compares random mating with this form of negative assortative mating for the complete range of starting frequencies. The negative assortative mating results in an excess of heterozygotes at equilibrium. Just as inbreeding can have a profound effect when combined with selection, so can negative assortative mating. Moreover, the excess of heterozygotes results in an accelerated rate of decrease in linkage disequilibrium, since the linkage phases are disrupted by recombination in double heterozygotes.

A second category of negative assortative mating is *pollen elimination*, which occurs when pollen of phenotype P_i fails to germinate on a stigma of phenotype P_i, but otherwise does not affect the fertility of the female plant. Under this system, the frequency of the fertilization P_i pollen $\times$ P_j female is $P_iP_j/(1 - P_i)$ and the frequency of the fertilization P_j pollen $\times$ P_i female is $P_iP_j/(1 - P_j)$. This model has quite different dynamics than the mass action model. In particular, pollen elimination results in only two stable equilibrium points, either $\hat{x} = 1/2$, $\hat{y} = 1/2$, $\hat{z} = 0$, or $\hat{x} = 0$, $\hat{y} = 1/2$ and $\hat{z} = 1/2$ (Workman 1964; Karlin and Feldman 1968). These equilibria contain only one heterozygote and the only permissible matings are homozygote $\times$ heterozygote. Half the progeny are again homozygous and half heterozygous, which perpetuates the equilibrium. The situation is analogous to that of the sexes in animals with sex chromosomes, in which there is a sort of "negative assortative mating" that perpetuates the sex ratio of 1:1.

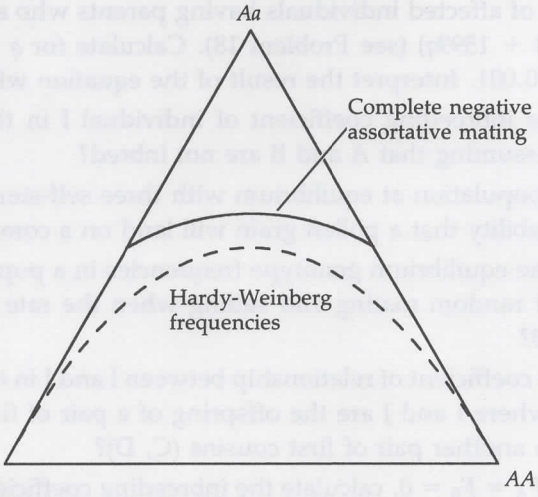


Figure 24. De Finetti diagram showing the genotype frequencies in a population with complete negative assortative mating. The population retains an excess of heterozygotes over the Hardy-Weinberg proportion.

PROBLEMS

1 What method could be used to calculate the inbreeding coefficient in organisms like bees in which males are the haploid result of development of unfertilized eggs?

2 In calculating F from pedigrees for X-linked genes, why are paths with two or more consecutive males not counted?

3 If a large population is maintained by self-fertilization in even-numbered generations and by random mating in odd-numbered generations, what happens to the inbreeding coefficient?

4 In the accompanying pedigree, what is the inbreeding coefficient of individual I when $F_A = 1/4$?

5 Calculate the inbreeding coefficient of the individual I in the accompanying pedigree.

6 Show that $p^2(1 - F) + pF = p^2 + pqF = p - (1 - F)pq$, when $q = 1 - p$.

7 With two alleles and $p = 1/2$, what are the expected genotype frequencies in a random mating population and among the offspring of first cousins? How great is the decrease in heterozygosity in the inbred population relative to the random mating population?

8 If the frequency of an autosomal recessive disorder is 1/1600 among unrelated parents, what is the expected frequency among the offspring of first cousins?

9 For a recessive allele at frequency q in a population in which 1% of the matings are between first cousins, but otherwise occur at random, the proportion of affected individuals having parents who are first cousins is $(1 + 15q)/(1 + 1599q)$ (see Problem 18). Calculate for $q = 0.1$, 0.05, 0.01, 0.005, and 0.001. Interpret the result of the equation when $q = 1$.

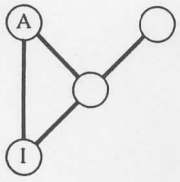
10 What is the inbreeding coefficient of individual I in the accompanying pedigree, assuming that A and B are not inbred?

11 In a plant population at equilibrium with three self-sterility alleles, what is the probability that a pollen grain will land on a compatible style?

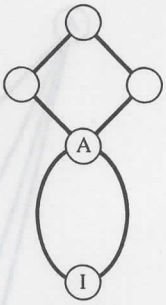
12 What are the equilibrium genotype frequencies in a population that practices mixed random mating and selfing when the rate of selfing is 20% and $p = 1/3$?

13 What is the coefficient of relationship between I and J in the accompanying pedigree, where I and J are the offspring of a pair of first cousins (A, B) mated with another pair of first cousins (C, D)?

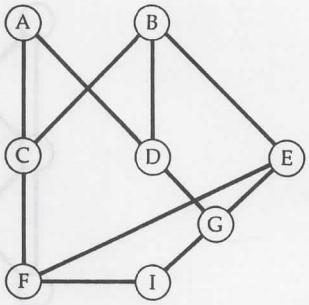
14 Assuming $F_A = F_B = 0$, calculate the inbreeding coefficient for each of the individuals C–I in the accompanying pedigree.



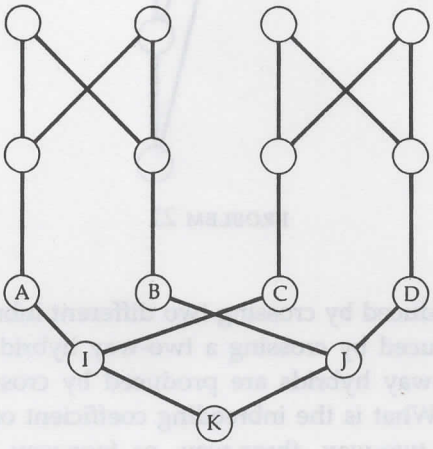
PROBLEM 4



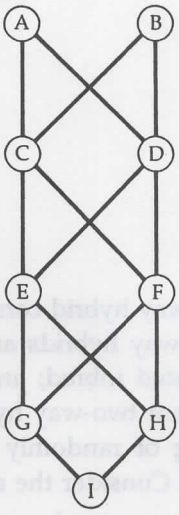
PROBLEM 5



PROBLEM 10



PROBLEM 13



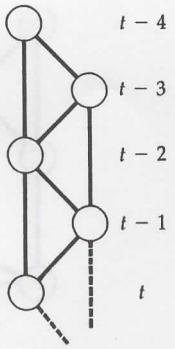
PROBLEM 14

15 For a gene with two alleles and $p = 0.3$, what are the expected genotype frequencies after 5 generations of sib mating? What are the expected genotype frequencies after one additional generation of random mating?

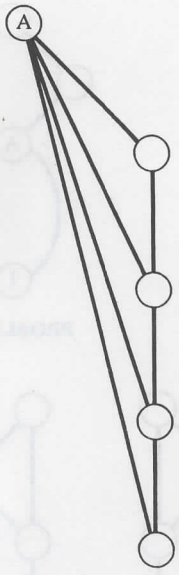
16 When random mating occurs with self-fertilization at rate S, show that $F_t = C + (S/2)^t (F_0 - C)$, where $C = S/(2 - S)$. For $S = 1/2$ and $F_0 = 0$, calculate F_1, F_2, F_3, and the equilibrium.

17 What are the analogous equations to those in Equation 5.32 for a system of n self-incompatibility alleles? What are the equilibrium frequencies?

18 For a recessive allele at frequency q in a population in which a proportion c of matings are between first cousins, but otherwise mating is random, show that the proportion of affected individuals having first-cousin parents is $c(1 + 15q)/(c + 16q - cq)$.



PROBLEM 20



PROBLEM 21

19 Two-way hybrid corn is produced by crossing two different inbred lines; three-way hybrids are produced by crossing a two-way hybrid with an unrelated inbred; and four-way hybrids are produced by crossing two different two-way hybrids. What is the inbreeding coefficient of the offspring of randomly mated two-way, three-way, or four-way hybrids? (Hint: Consider the allele frequencies in gametes.)

20 Derive a recursion equation for F_t for repeated parent-offspring mating (see pedigree), and calculate F_t for $t = 0$–5.

21 Derive a recursion equation for F_t for repeated backcrossing to a single noninbred individual A (see pedigree). Calculate F_t for $t = 0$–5 and the equilibrium value.

22 Derive a recursion equation for the pedigree in Problem 21 assuming that individual A has inbreeding coefficient F_A. What is the equilibrium in this case?

Chapter Six

Population Subdivision and Migration

Natural populations almost always display differences in allele and genotype frequencies from one geographic region to another. Such geographic population structure, previously ignored in our development of population genetic models, can have a profound effect on the evolutionary fate of a species. Almost any test of theoretical predictions would be invalidated by the presence of undetected population structure. For example, one of the simplest tests of genetic data from a population is the goodness-of-fit test to Hardy-Weinberg proportions, but this test is profoundly affected by including members of subpopulations with differing allele frequencies. The degree to which populations are subdivided can be quantified using Wright's F statistics, which have been applied to many populations and allow very informative comparisons. A wide variety of **migration models** have been devised to consider the effects of subdivision of populations into semi-isolated subpopulations, or **demes**, with various patterns of genetic exchange (gene flow) or migration among them. Combined effects of selection and migration have been examined, and some of the models predict geographic clines in allele frequency such as occur frequently in the natural world. A geographic **cline** in allele frequency is a steady increase or decrease in allele frequency across the geographical region. The cline models can be turned around to yield estimators of rates of migration using population samples. Finally, the effects of migration among semi-isolated demes can be an important factor in adaptive evolution when combined with the effects of natural selection and random genetic drift that occur within demes.

The Wahlund Effect with One Gene with Two Alleles

If two populations in Hardy-Weinberg genotypic proportions with different allele frequencies are pooled together, the pooled population will have an allele frequency equal to the average of those in the two populations. The effects of population pooling must be considered as potential complications to genetic analysis, because unrecognized subdivision may occur in samples from virtually all natural populations. One generation of random mating in the pooled population will result in Hardy-Weinberg proportions of genotypes, but the frequency of homozygous genotypes in the pooled population must be smaller than the average frequency of homozygotes in the two unmixed subpopulations. This reduction in homozygosity is called **Wahlund's principle**.

To reinforce the **Wahlund effect**, imagine two isolated subpopulations with allele a frequencies of q_1 and q_2. The average frequency of homozygous aa genotypes in the subpopulations is, from Table 1, $\bar{q}^2(1 - F) + \bar{q}F$, where $\bar{q}$ represents the average allele frequency of a over all subpopulations, and F represents the fixation index F_{ST}. Were the subpopulations to fuse into one, the allele frequency of a in the fused population would be $\bar{q}$. After one generation of random mating, in the absence of other forces, the frequency of homozygous aa genotypes is $\bar{q}^2$. Therefore, the average frequency of aa genotypes before population fusion is always greater than the frequency after fusion by the amount

$$\begin{aligned} \Delta R &= \bar{q}^2(1 - F) + \bar{q}F - \bar{q}^2 \\ &= \bar{q}^2F + \bar{q}F \\ &= \bar{q}(1 - \bar{q})F \end{aligned} \tag{6.1}$$

Table 1. The Wahlund effect: genotype frequencies in a subdivided population.

	AA	*Aa*	*aa*
Population 1	p_1^2	$2p_1q_1$	q_1^2
Population 2	p_2^2	$2p_2q_2$	q_2^2
Pooled (I)[a]	$(p_1^2 + p_2^2)/2$	$p_1q_1 + p_2q_2$	$(q_1^2 + q_2^2)/2$
Pooled (II)[a]	$\bar{p}^2(1 - F) + \bar{p}F$	$2\bar{p}\bar{q}(1 - F)$	$\bar{q}^2(1 - F) + \bar{q}F$

[a]The genotype frequencies in the pooled populations (I) and (II) are equivalent. Setting $(p_1^2 + p_2^2)/2 = \bar{p}^2(1 - F) + \bar{p}F$ and solving for F, one gets $F = \mathrm{Var}(p)/\bar{p}\bar{q}$.

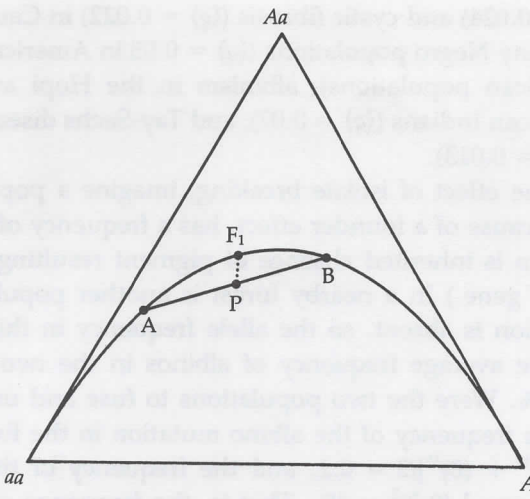


Figure 1. De Finetti diagram illustrating the Wahlund principle. Points A and B represent two subpopulations in Hardy-Weinberg equilibrium. The line connecting them represents the genotype frequencies resulting from a mixture of the genotypes from subpopulations A and B. Point P is a population composed of 50% from each of subpopulation A and B, and it clearly has a deficit of heterozygotes compared to the F_1 population.

A simple way to visualize the Wahlund effect on genotype frequencies is shown in the De Finetti diagram in Figure 1. The points labeled A and B represent two populations in Hardy-Weinberg equilibrium. The line connecting A and B represents the genotype frequencies of pooled populations containing any possible proportions of populations A and B. For example, the midpoint of the line, labeled P, represents a mixture that is half from population A and half from population B. The point P represents the genotype frequencies of a mixture of adults in a pooled population before any hybridization occurs. The effect of population mixing is immediately apparent from the fact that any point on such a line must lie below the Hardy-Weinberg parabola on a De Finetti diagram, and hence there is a deficit of heterozygotes in any mixture. After one generation of random mating, the genotype frequencies move to the point F_1, which is located on the parabola representing Hardy-Weinberg proportions of genotypes.

In human populations, the main effect of isolate breaking (fusion of populations) is to decrease the overall frequency of children born with genetic defects resulting from homozygous recessive genes, particularly harmful recessive genes with frequencies that are, for whatever reasons, relatively high in one of the populations. Examples of harmful recessive genes at high frequency in certain human populations include the alleles for α_1-antitrypsin

deficiency ($\langle q \rangle = 0.024$) and cystic fibrosis ($\langle q \rangle = 0.022$) in Caucasians; sickle-cell anemia in many Negro populations ($\langle q \rangle = 0.05$ in American blacks; $\langle q \rangle = 0.1$ in some African populations); albinism in the Hopi and some other Southwest American Indians ($\langle q \rangle = 0.07$); and Tay-Sachs disease among Ashkenazi Jews ($\langle q \rangle = 0.013$).

To illustrate the effect of isolate breaking, imagine a population of gray squirrels that, because of a founder effect, has a frequency of albinism equal to 16%. (Albinism is inherited absence of pigment resulting from a homozygous recessive gene.) In a nearby forest is another population in which the albino mutation is absent, so the allele frequency in this population is zero. Overall, the average frequency of albinos in the two populations is $(0.16 + 0)/2 = 8\%$. Were the two populations to fuse and undergo random mating, the allele frequency of the albino mutation in the fused population would be $[(0.16)^{1/2} + (0)^{1/2}]/2 = 0.2$, and the frequency of the homozygous genotype would equal $(0.2)^2 = 4\%$. That is, the frequency of albinos in the fused population is smaller than the average frequency among the individual isolates.

As indicated in Equation 6.1, the reduction in frequency of recessive homozygotes due to population fusion can be expressed in terms of the fixation index as $\bar{p}\bar{q}F$. The reduction in frequency can also be expressed in terms of the variance in allele frequency among subpopulations. The *variance* measures the degree of dispersion of a set of numbers from the mean, or how closely the individual numbers cluster around the mean. Consider two populations with allele frequencies q_1 and q_2, with $\bar{q} = (q_1 + q_2)/2$. Define a variable α that measures the difference between $\bar{q}$ and the allele frequency in each population, so that $q_1 = \bar{q} + \alpha$ and $q_2 = \bar{q} - \alpha$. The frequency of the *aa* genotype in a sample drawn with equal numbers from the two populations is

$$[(\bar{q} + \alpha)^2 + (\bar{q} - \alpha)^2]/2 = \bar{q}^2 + \alpha^2 \qquad \textbf{(6.2)}$$

Since α measures the deviation from the population mean allele frequency, α^2 is the sum of squared deviations, which is, by definition, the variance in allele frequency. Hence, in a subdivided population, the homozygote frequency is greater than the Hardy-Weinberg frequency in a hypothetical pooled population by an amount equal to the variance in allele frequency among the subpopulations. Letting this variance be denoted Var(p), the genotype frequencies are

	AA	*Aa*	*aa*
Subdivided	$\bar{p}^2 + \text{Var}(p)$	$2\bar{p}\bar{q} - 2\text{Var}(p)$	$\bar{q}^2 + \text{Var}(p)$
Pooled	$\bar{p}^2$	$2\bar{p}\bar{q}$	$\bar{q}^2$

Among any set of numbers, the variance can be calculated as the mean of

the squares (often called the mean square) minus the square of the mean. To take an extreme example, imagine two subpopulations with allele frequencies of 1 and 0, respectively. The mean square is $(1/2)[(1)^2 + (0)^2] = 0.5$ and the mean is $(1/2)[1 + 0] = 0.5$, so the variance Var(p) equals 0.5 [the mean square] $- (0.5)^2$ [the squared mean] $= 0.25$. The allele frequencies in the two populations in this example are as different as they can possibly be, so the maximum possible variance in allele frequency for two populations is 0.25. As another example, suppose the two allele frequencies are 0.75 and 0.25; then $\text{Var}(p) = (1/2)[(0.75)^2 + (0.25)^2] - [(1/2)(0.75 + 0.25)]^2 = 0.0625$. Note in this example that the deviations of the allele frequencies from the mean are half as large as in the previous example — 0.25 instead of 0.50 — but the variance is only one-quarter as large. This disproportionate reduction in the variance occurs because the variance depends on the square of the deviation of each allele frequency from the mean. As a final example, suppose the allele frequencies in the two populations are both 0.5. Then $\text{Var}(p) = (1/2)[(0.5)^2 + (0.5)^2] - [(1/2)(0.5 + 0.5)]^2 = 0$, which is simply the mathematical manner of stating that the populations have the same allele frequency (i.e., there is no variation). In all cases where there are just two alleles, $\text{Var}(p) = \text{Var}(1 - p) = \text{Var}(q)$.

The Wahlund Effect in Scandinavian Cod

A

The cod *Gadus morhua* is found throughout the shallow waters of coastal areas, and in 1965 Knud Sick undertook a survey of the hemoglobin polymorphism of this species in the Scandinavian waters from the northern tip of Jutland into the eastern Baltic Sea (Figure 2). The sampling area was somewhat arbitrarily split into three regions, with the central area called Region 3. Observed and expected genotype frequencies based on Hardy-Weinberg proportions were as follows, along with χ^2 and P values for goodness-of-fit:

	Hb^1/Hb^1	Hb^1/Hb^2	Hb^2/Hb^2	χ^2	P
Region 1	511 (521.7)	685 (663.6)	200 (210.7)	1.453	0.23
Region 2	3 (0.5)	29 (33.9)	538 (535.5)	0.179	0.90
Region 3	418 (180.8)	749 (1223.3)	2306 (2068.8)	522.117	0

It is clear that the central populations (Region 3) have an excess of homozygotes, leading to the suspicion that these populations may represent a mixture of fish from the two adjacent regions. Region 3 is actually composed of a set of samples with quite heterogeneous allele frequencies, making it inappropriate to test the fit to Hardy-Weinberg proportions. The data were actually collected from 50 samples throughout the range, and the data can be used to determine whether the departure from Hardy-Weinberg proportions in Region 3 is due to the Wahlund effect. In Figure 3, the allele frequency in each sample is plotted against the value of the χ^2 that tests the goodness-of-fit to Hardy-Weinberg proportions. In agreement with the model, in the eastern population, where all the samples had allele frequencies of around 3% Hb^1, the χ^2 are

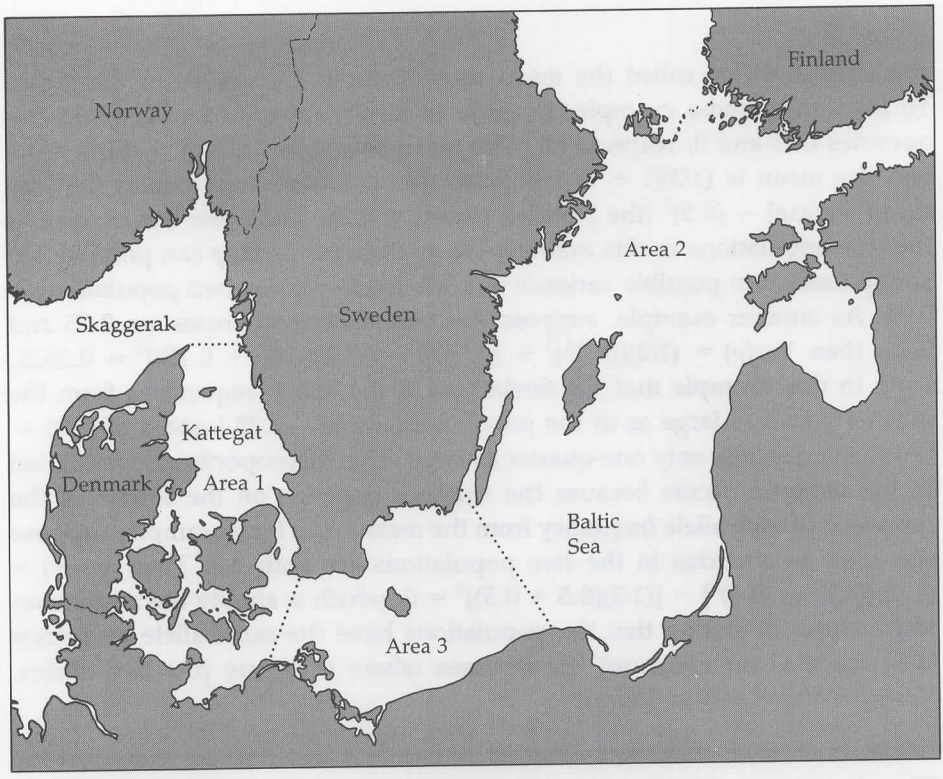


Figure 2. Map of Scandinavia showing the Skaggerak, the Kattegat, and the Baltic and the arbitrary subdivision into three areas used by Sick (1965) in a study of cod.

small (good fit to Hardy-Weinberg), and so also in the western samples, where the allele frequency was consistently around 61% Hb^1. The intermediate populations have a range of allele frequencies, possibly reflecting a range of mixtures of the two populations, and the χ^2 is greatest in the mixtures closest to 50:50.

The Wahlund principle states that heterozygote frequencies for a two-allele gene should be $2pq - 2\text{Var}(p)$, where $\text{Var}(p)$ is the variance in allele frequencies among subpopulations. $\text{Var}(p)$ can be estimated from the hemoglobin data by assuming that any sample represents a mixture of two populations with allele frequencies 0.03 and 0.61. From the observed allele frequency in a sample, the proportions of the two originating populations in the mixture can be determined from the formula

$$x = \frac{p_1 - p_3}{p_1 - p_2}$$

where $p_1 = 0.61$, $p_2 = 0.03$, and p_3 is the allele frequency in any particular sample from Region 3. The quantity denoted x is the fraction of the Region 3 population that came from the region with allele frequency 0.03. Knowing x, one can calculate $\text{Var}(p)$ as $xp_2^2 + (1 - x)p_1^2 - [xp_2 + (1 - x)p_1]^2$. When the estimate of 2 $\text{Var}(p)$ for each sample is plotted against the expected-minus-observed heterozygote frequency, a linear relation is expected. Figure 4 shows a reasonable fit to this expectation.

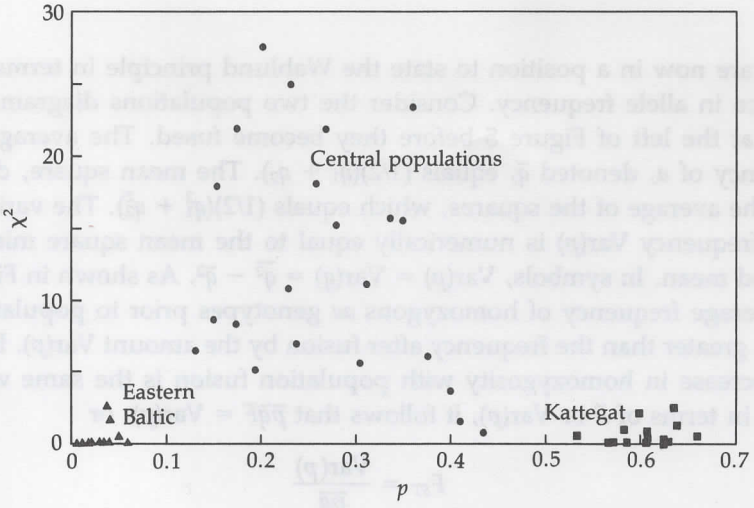


Figure 3. Chi-square values for the departure from panmixia in samples of cod taken from the Eastern Baltic (where $p < 0.1$), from central populations (where $p \approx 0.3$), and from the Kattegat (where $p > 0.5$). Note that the central populations differ greatly from Hardy-Weinberg proportions, suggesting that the sample is a mixture (there is a deficit of heterozygotes in the central populations).

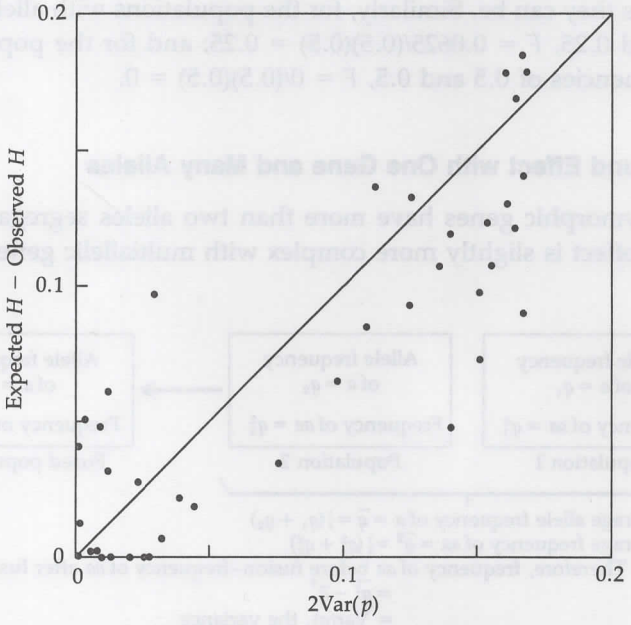


Figure 4. According to the Wahlund principle, the expected frequency of heterozygotes is $2pq - 2\text{Var}(p)$. For each of the 50 samples taken in the central region (Box A), the difference between the observed and expected heterozygosity is plotted against the variance in allele frequency for that sample estimated from the degree of admixture as described in the text. (Data from Sick 1965.)

We are now in a position to state the Wahlund principle in terms of the variance in allele frequency. Consider the two populations diagrammed in boxes at the left of Figure 5 before they become fused. The average allele frequency of a, denoted $\bar{q}$, equals $(1/2)(q_1 + q_2)$. The mean square, denoted $\overline{q^2}$, is the average of the squares, which equals $(1/2)(q_1^2 + q_2^2)$. The variance in allele frequency Var(p) is numerically equal to the mean square minus the squared mean. In symbols, $\text{Var}(p) = \text{Var}(q) = \overline{q^2} - \bar{q}^2$. As shown in Figure 5, the average frequency of homozygous *aa* genotypes prior to population fusion is greater than the frequency after fusion by the amount Var(p). Because the decrease in homozygosity with population fusion is the same whether stated in terms of F or Var(p), it follows that $\bar{p}\bar{q}F = \text{Var}(p)$, or

$$F_{ST} = \frac{\text{Var}(p)}{\bar{p}\bar{q}} \tag{6.3}$$

Equation 6.3 provides a convenient method of estimating F from allele frequency data. For example, in the three pairs of hypothetical populations discussed earlier, $\bar{p} = \bar{q} = 0.5$ in all cases. For the populations with allele frequencies of 1 and 0, $F = \text{Var}(p)/\bar{p}\bar{q} = 0.25/(0.5)(0.5) = 1$, which is simply a statement in terms of the fixation index F_{ST} that the populations are as different as they can be. Similarly, for the populations with allele frequencies of 0.75 and 0.25, $F = 0.0625/(0.5)(0.5) = 0.25$; and for the populations with allele frequencies of 0.5 and 0.5, $F = 0/(0.5)(0.5) = 0$.

The Wahlund Effect with One Gene and Many Alleles

Many polymorphic genes have more than two alleles segregating, and the Wahlund effect is slightly more complex with multiallelic genes. In the case

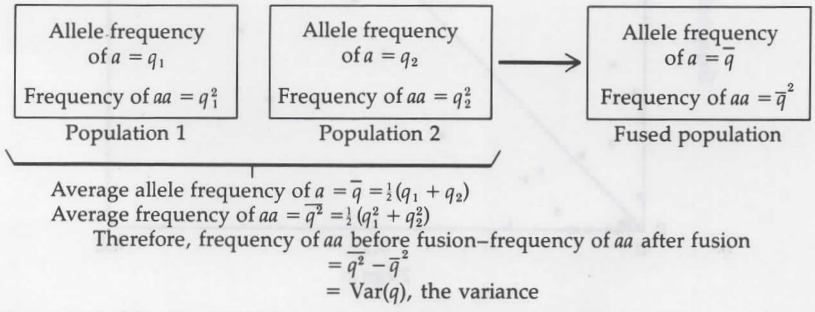


Figure 5. Illustration of Wahlund's principle (isolate breaking). The frequency of homozygous recessives upon population fusion and random mating is less than the average frequency before fusion by an amount equal to the variance in allele frequency.

of three alleles, A_1, A_2, and A_3, with respective frequencies p_i, q_i, and r_i in the ith population, the genotype frequencies within any population are: A_1A_1, p_i^2; A_1A_2, $2p_iq_i$; A_2A_2, q_i^2; and so on. In the pooled population, the frequencies of the three alleles are $\bar{p}$, $\bar{q}$, and $\bar{r}$. The genotype frequencies in the pooled population are

Genotype	Frequency	
A_1A_1	$\bar{p}^2 + \mathrm{Var}(A_1)$	
A_1A_2	$2\bar{p}\bar{q} + 2\mathrm{Cov}(A_1,A_2)$	
A_1A_3	$2\bar{p}\bar{r} + 2\mathrm{Cov}(A_1,A_3)$	**(6.4)**
A_2A_2	$\bar{q}^2 + \mathrm{Var}(A_2)$	
A_2A_3	$2\bar{q}\bar{r} + 2\mathrm{Cov}(A_2,A_3)$	
A_3A_3	$\bar{r}^2 + \mathrm{Var}(A_3)$	

where $\mathrm{Var}(A_i)$ is the variance in the frequency of allele A_i across populations (Nei 1965; Li 1969). With more than two alleles, the variances of the allele frequencies among populations can differ. The Cov terms are the covariance across populations. For example, $\mathrm{Cov}(A_1,A_2)$ is the covariance of alleles A_1 and A_2 across populations, which is calculated as $\Sigma_i p_i q_i - \bar{p}\bar{q}$. Analogous calculations give the covariances for the other combinations of alleles. Covariances can be positive or negative, so that the frequency of a particular heterozygote in a subdivided population may be either greater or less than the Hardy-Weinberg value based on the average allele frequencies. On the other hand, the total heterozygosity is always depressed below the panmictic proportions in a subdivided population, because all of the covariances cannot be positive. This is shown graphically in Figures 6 and 7. Figure 6 shows the total heterozygote frequency for a three-allele panmictic population, which is

$$H = 1 - \bar{p}^2 - \bar{q}^2 - \bar{r}^2 \qquad \textbf{(6.5)}$$

The points A and B in Figure 6 represent two subpopulations that differ in allele frequency. The chord that connects the two points represents the heterozygosity of various mixtures of A and B. Since the surface curves in such a way as always to lie above any chord connecting two points on the surface, a mixture of populations always has a lower total heterozygosity than does the F_1 panmictic population. Figure 7 shows the scallop-shaped surface of frequency of A_1A_2 heterozygotes as a function of the frequencies of alleles A_1 and A_2. When one population has a low frequency of both alleles and another population has a high frequency of both alleles, then the allele frequencies have a positive covariance. The line that connects two such populations (labeled A and B in Figure 6) lies above the surface, meaning

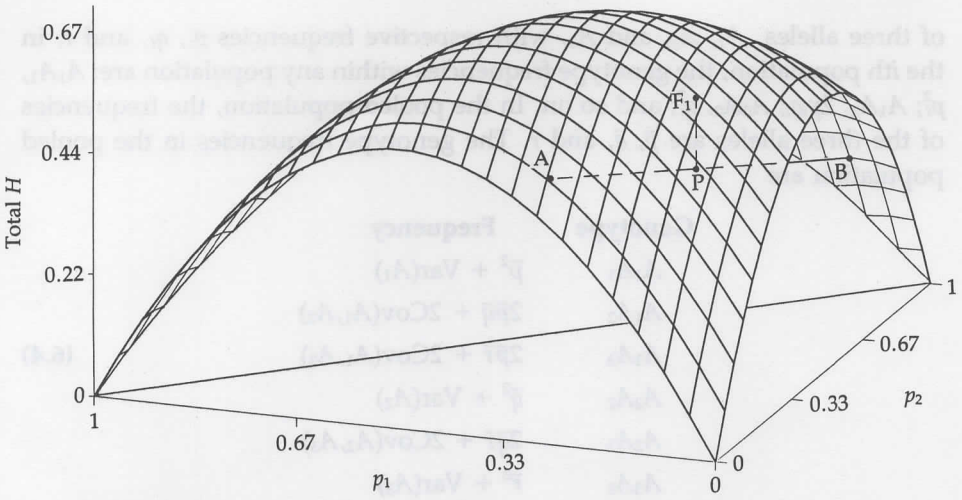


Figure 6. The Wahlund principle with three alleles holds with respect to total heterozygosity. The curved surface is the total heterozygosity in a Hardy-Weinberg population ($H = 1 - \Sigma p_i^2$), plotted here against the frequencies of two of the alleles (p_1 and p_2). Points A and B are subpopulations in Hardy-Weinberg equilibrium, and the line connecting them represents pooled populations (mixtures of A and B). The pooled population P has a deficit of heterozygotes compared to the F_1 population.

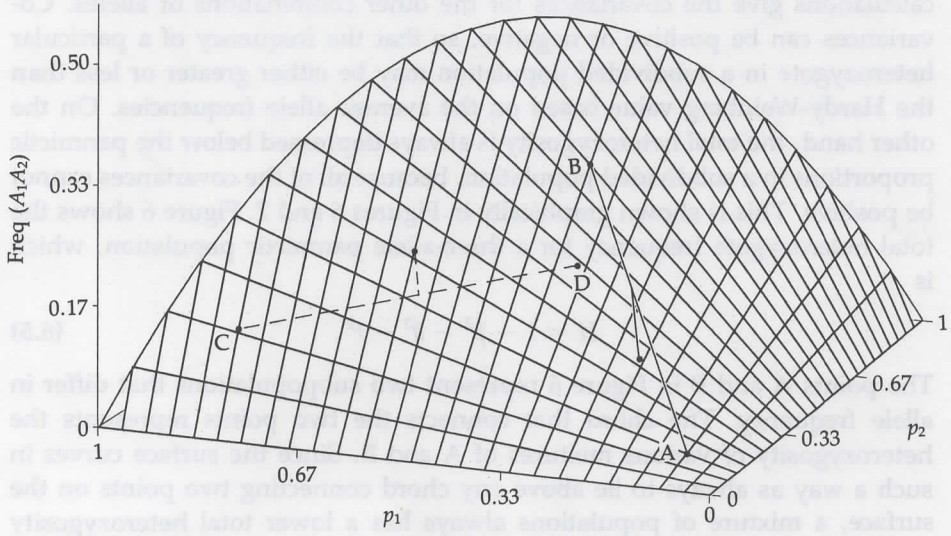


Figure 7. Illustration of why there is not always a deficit of heterozygotes for all pairs of alleles when two populations differing in allele frequency at a multiallelic locus are pooled. The line connecting A and B lies above the scallop-shaped surface that represents the frequency of the A_1A_2 heterozygote in a panmictic population, and the frequency of this heterozygote is therefore greater in the pooled population than it will be in the F_1. The line connecting C and D lies beneath the surface, so a pooling of these populations would have a frequency of A_1A_2 less than in the F_1.

that the pooled population has a higher frequency of the A_1A_2 heterozygotes than in a panmictic population with allele frequencies $\bar{p}$ and $\bar{q}$. On the other hand, when one population has a high frequency of allele A_1 and a low frequency of allele A_2, while another population has a low frequency of A_1 and a high frequency of A_2, then the allele frequencies have a negative covariance across populations. In this case, the chord that connects two such populations (labeled C and D on the figure) lies beneath the surface, so — just as in the case of total heterozygosity — when there is negative covariance of allele frequencies, the mixed population has a deficit of a particular heterozygote frequency compared to the panmictic proportions.

The fact that not all heterozygotes show a deficit in structured populations points out an important distinction between subdivision and inbreeding. With inbreeding, the reduction in heterozygote frequencies is the same for all genes, including multiallelic genes, while with population subdivision, the frequencies of heterozygotes differ from panmictic proportions in a way that depends on the covariances of allele frequencies across populations.

The Wahlund Effect with Two Genes

When Hardy-Weinberg populations that differ in two-locus gametic frequencies are fused, the resulting population is likely to manifest linkage disequilibrium (Nei and Li 1973). This is most easily seen by an example. The table below shows gametic frequencies for two populations, each of which is in linkage equilibrium, and a 50:50 mixture of the two.

	Gametic frequencies			
	AB	***Ab***	***aB***	***ab***
Population X	x_1	x_2	x_3	x_2x_3/x_1
Population Y	y_1	y_2	y_3	y_2y_3/y_1
Mixture	$(x_1 + y_1)/2$	$(x_2 + y_2)/2$	$(x_3 + y_3)/2$	$(x_2x_3/x_1+y_2y_3/y_1)/2$

The gametic frequencies of the mixture are simply the average of the frequencies in the two populations. Using the formula $D = x_1x_4 - x_2x_3$, we find that the value of D is zero in each of populations X and Y, but the mixture has

$$D = (1/4)\left(\frac{x_1y_2y_3}{y_1} - x_2y_3 - x_3y_2 + \frac{x_2x_3y_1}{x_1}\right) \qquad \textbf{(6.6)}$$

If the gametic frequencies in the two original populations are different, then D in the pooled population will generally be nonzero. This principle is particularly easy to see in the extreme case of $x_1 = 1$, $x_2 = 0$, $x_3 = 0$, and $x_4 = 0$ for one population; and $y_1 = 0$, $y_2 = 0$, $y_3 = 0$, and $y_4 = 1$ for the second population. An equal mixture of these populations has gametic fre-

quencies $z_1 = 0.5$, $z_2 = 0$, $z_3 = 0$, and $z_4 = 0.5$, giving $D = 0.25$—the maximum possible value. If the other two gametes had been present (i.e., $x_2 = 1$ and $y_3 = 1$) then the pooled population would have $D = -0.25$. Hence, population pooling can lead to either positive or negative linkage disequilibrium. Figure 8 shows diagrammatically how mixtures of populations can have nonzero linkage disequilibrium, even when the component subpopulations have $D = 0$. Note that the surface plotted for $D = 0$ curves "in both directions." This implies that there should be some special pairs of populations that have different gametic frequencies, but $D = 0$ in the pooled population as well. Indeed, examples are not hard to find: $x_1 = 0.6$, $x_2 = 0.6$, $x_3 = 0.4$, and $x_4 = 0.4$, for one population; and $y_1 = 0.4$, $y_2 = 0.4$, $y_3 = 0.6$, and $y_4 = 0.6$ for the second population. Both of these populations are in linkage equilibrium, and the pooled population has $z_1 = z_2 = z_3 = z_4 = 0.5$, so $D = 0$ in the pooled population as well. By a straightforward extension of this reasoning, migration among a series of semi-isolated populations can result in a cline in linkage disequilibrium (Feldman and Christiansen 1975).

So far we have considered only gametic frequencies, while the pooled population will also have genotypic frequencies that differ from the two-locus Hardy-Weinberg proportions. As in the case of multiallelic genes, whether a particular heterozygous genotype is present in excess or deficit

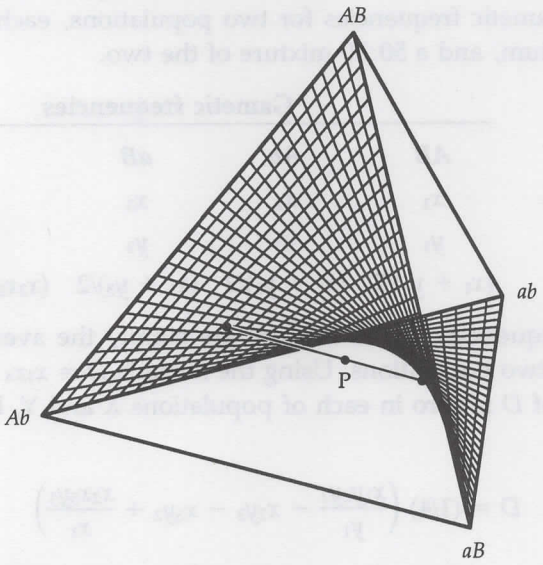


Figure 8. The Wahlund principle illustrated graphically for the case of two loci. The surface represents $D = 0$, and a combination of any two populations lying on the surface can yield a pooled population (P) with gametic frequencies no longer in linkage equilibrium.

depends on the covariances of gametic frequencies across populations. The full analysis of two-locus gametic and genotype frequencies in subdivided populations is very complex and involves digenic descent measures (Weir and Cockerham 1988).

WRIGHT'S *F* STATISTICS

Population subdivision entails an inbreeding-like effect in terms of excess homozygosity; it is therefore convenient to measure this effect in terms of the decrease in the proportion of heterozygous genotypes, much as the effects of nonrandom mating were measured in Chapter 5. A subdivided population, however, has three distinct levels of complexity: individual organisms (I), subpopulations (S), and the total population (T). To avoid potential confusion of these various levels, it is convenient to adopt the terminology below. We will also allow the possibility of nonrandom mating within subpopulations.

H_I = the heterozygosity of an individual in a subpopulation

H_S = the expected heterozygosity of an individual in an equivalent random mating subpopulation

H_T = the expected heterozygosity of an individual in an equivalent random mating total population

H_I can be interpreted as the average heterozygosity of all the genes in an individual or as the probability of heterozygosity of any one gene. H_I is the observed heterozygosity averaged across subpopulations. H_S represents the level of heterozygosity that would be found in a subpopulation if the subpopulation were undergoing random mating; therefore, H_S always equals $2p_iq_i$ for a subpopulation with allele frequency p_i. The quantity H_T represents what the heterozygosity would be if all subpopulations were pooled together and mated randomly; if the average allele frequency among subpopulations is denoted p_0, then $H_T = 2p_0q_0$.

Wright developed his F statistics in terms of one locus with two alleles, and treated multiallelic data by keeping the most common allele and pooling the other alleles. An explicit multiallelic formulation was presented by Nei (1973). With multiple alleles the definitions of H_I, H_S, and H_T remain the same, but it is easier to see what they measure by inspecting the formulas. H_I is the observed heterozygosity averaged over subpopulations. If H_i is the heterozygosity in subpopulation i, then for the case of k subpopulations,

$$H_I = \sum_{i=1}^{k} H_i/k \qquad (6.7)$$

If $p_{i,s}$ is the frequency of the ith allele in subpopulation s, and H_S is the expected Hardy-Weinberg heterozygosity in subpopulation s, then

$$H_S = 1 - \sum_{i=1}^{k} p_{i,s}^2 \tag{6.8}$$

where the summation is over all i alleles. $\overline{H}_S$ is the average of H_S taken over subpopulations. Finally, if $\overline{p}_i$ is the frequency of allele i averaged over the subpopulations, then

$$H_T = 1 - \sum_{i=1}^{k} \overline{p}_i^2 \tag{6.9}$$

The inbreeding coefficient defined in Chapter 5 measures the reduction in heterozygosity of an individual due to nonrandom mating within its subpopulation. To avoid confusion, we will call this inbreeding coefficient F_{IS} (rather than simply F as in Chapter 5). Thus,

$$F_{IS} = \frac{\overline{H}_S - H_I}{\overline{H}_S} \tag{6.10}$$

The effects of population subdivision are measured by a quantity called the **fixation index** (symbolized F_{ST}), which is the reduction in heterozygosity of a subpopulation due to random genetic drift. Thus,

$$F_{ST} = \frac{H_T - \overline{H}_S}{H_T} \tag{6.11}$$

F_{ST} is always greater than (or equal to) zero, because the Wahlund effect assures that $H_T > \overline{H}_S$. If all subpopulations are in Hardy-Weinberg equilibrium with the same allele frequencies, $F_{ST} = 0$. A graphical demonstration of this appears in Box B.

B The Wahlund Effect Assures that $H_T > \overline{H}_S$

Figure 9 shows the frequency of heterozygotes ($2pq$) plotted against the allele frequency p for a gene with two alleles. Two populations — A and B — are indicated, as are their Hardy-Weinberg heterozygosities, which are called $H_{S\text{A}}$ and $H_{S\text{B}}$. The mean of the two values is $\overline{H}_S$. The point midway between A and B represents a population whose allele frequency is that of the pooled population, which is labeled $\overline{p}$. The heterozygosity of a Hardy-Weinberg population with this allele frequency is, by definition, H_T. Note that H_T must be greater than $\overline{H}_S$. This will always be the case when the heterozygosity function is concave downward, which is guaranteed by a mathematical result known as the Jensen inequality (Karlin and Taylor 1975, p. 249). The Jensen inequality states that, for a function that is concave downward, the function of an expectation (in this case H_T) is greater than the expectation of the function (in this case $\overline{H}_S$). For our purposes,

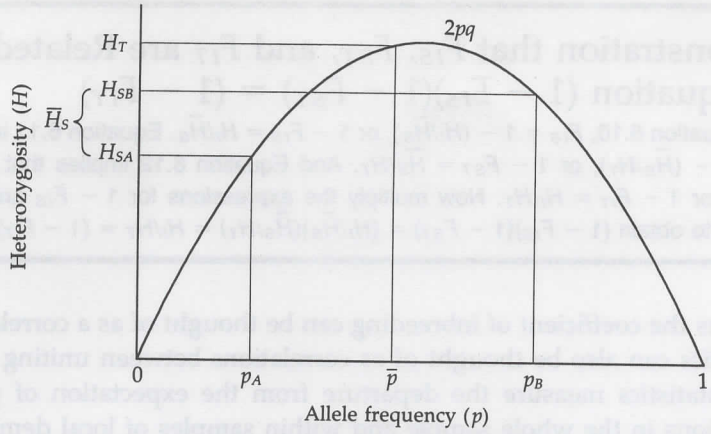


Figure 9. Two subpopulations with allele frequencies p_A and p_B have Hardy-Weinberg heterozygosities H_{SA} and H_{SB}, respectively. From the definition $\overline{H}_S$ is the average of these two, or $\overline{H}_S = (H_{SA} + H_{SB})/2$. A population with allele frequency $\overline{p} = (p_A + p_B)/2$ has Hardy-Weinberg heterozygosity H_T. The geometry of these relationships dictates that $H_T \geq \overline{H}_S$.

the result is readily apparent by studying the graph. In the case of multiple alleles, the total heterozygosity function is always concave downward, but heterozygosity for individual alleles is concave upward in certain planes.

The overall inbreeding coefficient of an individual (denoted F_{IT}) includes a contribution due to actual nonrandom mating within subpopulations (F_{IS}) and another contribution due to the subdivision itself (F_{ST}). The quantity F_{IT} measures the reduction in heterozygosity of an individual relative to the total population, so

$$F_{IT} = \frac{H_T - H_I}{H_T} \qquad \textbf{(6.12)}$$

The hierarchical F statistics defined in Equations 6.10, 6.11 and 6.12 are all types of "inbreeding coefficients," but they differ according to the reference populations. F_{IS} is concerned with inbreeding in individuals (I), relative to the subpopulation (S) to which they belong; F_{ST} is concerned with inbreeding in subpopulations (S), relative to the total population (T) of which they are a part; and F_{IT} is concerned with inbreeding in individuals (I), relative to the total population (T). F_{IT} is the most inclusive measure of inbreeding in that it takes into account both the effects of nonrandom mating within subpopulations (F_{IS}) and the effects of population subdivision (F_{ST}). The mathematical relation between the three types of inbreeding coefficient is demonstrated in Box C.

C Demonstration that F_{IS}, F_{ST}, and F_{IT} are Related by the Equation $(1 - F_{IS})(1 - F_{ST}) = (1 - F_{IT})$

From Equation 6.10, $F_{IS} = 1 - (H_I/\bar{H}_S)$, or $1 - F_{IS} = H_I/\bar{H}_S$. Equation 6.11 implies that $F_{ST} = 1 - (\bar{H}_S/H_T)$, or $1 - F_{ST} = \bar{H}_S/H_T$. And Equation 6.12 implies that $F_{IT} = 1 - (H_I/H_T)$, or $1 - F_{IT} = H_I/H_T$. Now multiply the expressions for $1 - F_{IS}$ and $1 - F_{ST}$ together to obtain $(1 - F_{IS})(1 - F_{ST}) = (H_I/\bar{H}_S)(\bar{H}_S/H_T) = H_I/H_T = (1 - F_{IT})$.

Just as the coefficient of inbreeding can be thought of as a correlation, the F statistics can also be thought of as correlations between uniting gametes. The F statistics measure the departure from the expectation of panmictic proportions in the whole sample and within samples of local demes. The F statistics are equivalent to correlation coefficients that measure the correlations between uniting gametes under various hypothetical and actual sampling schemes. F_{IT} is the correlation of uniting gametes relative to gametes drawn at random from the entire population. F_{IS} is the correlation of uniting gametes relative to gametes drawn at random from within a subpopulation (averaged over subpopulations), and F_{ST} is the correlation of gametes within subpopulations relative to gametes drawn at random from the entire population.

Equations 6.10–6.12 can be applied to data of Levin (1978) on natural subpopulations of the flowering plant *Phlox cuspidata* in Texas. Altogether, 43 subpopulations were studied with respect to the phosphoglucomutase-2 (*Pgm-2*) gene, and two alleles ($Pgm\text{-}2^a$ and $Pgm\text{-}2^b$) were found. In 40 subpopulations, the $Pgm\text{-}2^b$ allele was fixed; in the remaining three subpopulations, the allele frequencies of $Pgm\text{-}2^b$ were estimated as 0.49, 0.83, and 0.91, respectively. In the three polymorphic populations, the observed frequencies of heterozygotes were 0.17, 0.06, and 0.06, respectively.

The individual heterozygosity (H_I) can be estimated from the observed data as the average heterozygosity of individuals within subpopulations; thus $\langle H_I \rangle = [(40)(0) + 0.17 + 0.06 + 0.06]/43 = 0.0067$, where the first term comes from the 40 subpopulations in which the observed heterozygosity was zero. The subpopulation heterozygosity ($\bar{H}_S$) can be estimated from the estimated allele frequencies within subpopulations and the expected frequency of heterozygotes under the supposition of random mating; thus $\langle \bar{H}_S \rangle = [(40)(0) + 2(0.49)(0.51) + 2(0.83)(0.17) + 2(0.91)(0.09)]/43 = 0.0220$. Here again the first term comes from the 40 fixed subpopulations. Finally, the total heterozygosity (H_T) can be estimated from the average allele frequency in the combined subpopulations and the Hardy-Weinberg expectation of heterozygosity with random mating; the average allele frequency of the $Pgm\text{-}2^b$ allele is $[(40)(1) + 0.49 + 0.83 + 0.91]/43 = 0.9821$, so $\langle H_T \rangle = 2(0.9821)(0.0179) = 0.0352$.

Now the various F statistics can be estimated from Equations 6.10–6.12:

$\langle F_{IS} \rangle = (\langle \overline{H}_S \rangle - \langle H_I \rangle)/\langle \overline{H}_S \rangle = (0.0220 - 0.0067)/(0.0220) = 0.70$. This is the estimated inbreeding coefficient due to nonrandom mating within subpopulations. The fixation index F_{ST} is estimated as $\langle F_{ST} \rangle = (\langle H_T \rangle - \langle \overline{H}_S \rangle)/\langle H_T \rangle = (0.0352 - 0.0220)/(0.0352) = 0.36$; this is the amount of inbreeding due solely to population subdivision. Finally, the total inbreeding coefficient is estimated as $\langle F_{IT} \rangle = (\langle H_T \rangle - \langle H_I \rangle)/\langle H_T \rangle = (0.0352 - 0.0067)/(0.0352) = 0.81$; this is the amount of inbreeding due to the combined effects of nonrandom mating within subpopulations and to random genetic drift among subpopulations. Note that the total inbreeding coefficient is quite large (as expected from the substantial fraction of self-fertilization in this organism), but an appreciable fraction of the total inbreeding is attributable to subdivision and random genetic drift.

The method of estimating the F statistics by replacing the parameters in Equations 6.10–6.12 with their observed or estimated values is not necessarily the best, particularly with small samples. Ideally, estimates of the F statistics should correct for the effects of sampling a limited number of subpopulations, as well as for the effects of sampling a limited number of individuals per subpopulation. Methods for making these corrections have been developed, but they are quite complex and raise additional issues. For an excellent discussion, see Weir and Cockerham (1984). Important issues are also addressed in Wright (1978, p. 86), Curie-Cohen (1982), Nei and Chesser (1983), and Nei (1986). When applied to the *Pgm* data from *Phlox cuspidata*, for example, the more sophisticated method of Weir and Cockerham gives $F_{IS} = 0.70$, $F_{ST} = 0.36$, and $F_{IT} = 0.81$, in agreement with the uncorrected procedure in the first two decimal places. For purposes of illustration, the uncorrected estimation procedure used above is adequate.

Estimating F Statistics from Population Samples

D

Levin (1978) also studied the glutamate oxalacetate transaminase-2 (*Got*-2) gene in the 43 subpopulations of *Phlox cuspidata* discussed earlier. Three alleles were found: *Got*-2^a, *Got*-2^b and *Got*-2^c. A total of 39 subpopulations were monomorphic for *Got*-2^b. One subpopulation contained *Got*-2^a and *Got*-2^b with respective allele frequencies 0.37 and 0.63, and the observed frequency of heterozygotes was 0.17. Three subpopulations contained only *Got*-2^b and *Got*-2^c; the allele frequencies of *Got*-2^b in these subpopulations were 0.87, 0.91, and 0.82, respectively, and the observed frequencies of heterozygotes were 0.09, 0.06, and 0.09, respectively. From these samples, estimates of H_I, H_S, H_T, F_{IS}, F_{IT} and F_{ST} are readily obtained. The formal analysis requires correction for sampling error, which we ignore here.

$$\langle H_I \rangle = [(39)(0) + 0.17 + 0.09 + 0.06 + 0.09]/43$$
$$= 0.00953$$
$$\langle \overline{H}_S \rangle = [(39)(0) + 2(0.37)(0.63) + 2(0.87)(0.13) + 2(0.91)(0.09) + 2(0.82)(0.18)]/43$$
$$= 0.02678$$

The pooled allele frequency of *Got-2*^{a} is [(39)(0) + 0.37 + 3(0)]/43 = 0.0086; *Got-2*^{b} is [(39)(1) + 0.63 + 0.87 + 0.91 + 0.82]/43 = 0.9821, so that of *Got-2*^{c} is 1 − 0.0086 − 0.9821 = 0.0093. From the Hardy-Weinberg principle for multiple alleles (Chapter 1),

$$\langle H_T \rangle = 2(0.0086)(0.9821) + 2(0.0086)(0.0093) + 2(0.9821)(0.0093) = 0.03532$$

Thus $\langle F_{IS} \rangle = 0.64$, $\langle F_{ST} \rangle = 0.24$, and $\langle F_{IT} \rangle = 0.73$. As you may have noted, the F statistics vary somewhat from gene to gene; the range of $\langle F_{IS} \rangle$ for five genes studied in *Phlox cuspidata* was 0.60 to 0.77 (average 0.67), and that of $\langle F_{ST} \rangle$ was 0.17 to 0.86 (average 0.41).

Except for plants that have a high frequency of self-fertilization or for certain insects that regularly undergo parent–offspring or brother–sister mating, values of F_{IS} in most natural populations are typically close to zero, which indicates random mating within subpopulations. (Within very small subpopulations, there may be mating between relatives — but as long as the mating between relatives occurs no more often than would be expected by chance in a subpopulation of that size, F_{IS} remains zero.) For example, Crow and Mange (1965) studied a South Dakota population of Hutterites, a religious group that is relatively isolated from surrounding communities because of their religious beliefs. In this population, F_{ST} was estimated as 0.04 (a relatively large value for a human population), yet F_{IS} was essentially zero. For many natural populations, therefore, particularly in animals and outcrossing plants, it is reasonable to assume $F_{IS} = 0$. This assumption is equivalent to the assumption of random mating within subpopulations. Furthermore, when $F_{IS} = 0$, then $F_{ST} = F_{IT}$, and there is no longer need for the subscripts. In the context of multiallelic loci, F_{ST} is denoted G_{ST} (Nei 1973).

E Cockerham's Partitioning of F by Analysis of Variance

A convenient way to put the F statistics into a familiar context for hypothesis testing was developed by Cockerham (1969, 1973). Consider a two-allele polymorphism with alleles A and a with frequencies p and $q = 1 - p$, respectively. The alleles are arbitrarily assigned the values $x = 1$ for allele A and $x = 0$ for allele a. The expectation of x is simply the allele frequency p and the total variance is $p(1 - p)$. The variance of x (σ^2) can be partitioned as follows:

$$x_{ijk} = p + a_k + b_{ik} + w_{ijk} \qquad \textbf{(6.13)}$$

where x_{ijk} is the value (1 or 0) of the jth allele in the ith individual in the kth subpopulation. The effect denoted a_k is for subpopulations, b_{ik} for the individuals within subpopulations, and w_{ijk} for alleles within individuals (inbreeding). In the terminology of analysis of variance, these are all random effects with associated variances σ_a^2, σ_b^2, and σ_w^2, respectively.

The hierarchical F statistics are related to these components of variance as follows:

$$\begin{aligned} \sigma_w^2 &= (1 - F_{IT})\sigma^2 \\ \sigma_b^2 &= (F_{IT} - F_{ST})\sigma^2 \\ \sigma_a^2 &= F_{ST}\sigma^2 \end{aligned} \tag{6.14}$$

These are readily solved to give the F statistics in terms of variance components.

$$F_{IT} = \frac{\sigma_a^2 + \sigma_b^2}{\sigma^2}$$

$$F_{ST} = \frac{\sigma_a^2}{\sigma^2} \tag{6.15}$$

$$F_{IS} = \frac{\sigma_b^2}{\sigma_b^2 + \sigma_w^2}$$

Cockerham showed how to estimate these components using analysis of variance, as depicted in Table 2. By relating F statistics to analysis of variance, it is also possible to add levels of population subdivision, and to extract their effects as additional components of variation.

Table 2. Analysis of variance of allele frequencies in a subdivided population.

Source	df	Mean squares	Expected mean squares
Among subpopulations	$M - 1$	MS_a	$\sigma_w^2 + 2\sigma_b^2 + 2N\sigma_a^2$
Among individuals within subpopulations	$M(N - 1)$	MS_b	$\sigma_w^2 + 2\sigma_b^2$
Within individuals	N	MS_w	σ_w^2

Equal sample size within each subpopulation is assumed. (From Cockerham 1969.)

The estimation of F statistics becomes more complicated with multiple alleles and multiple genes, and there is no single method that is universally the best. Long (1986) has presented a method that is a nice extension of Cockerham's approach. Rather than indexing the two alleles as 1 and 0, Long (1986) assigns a vector of length $k - 1$ for each allele at a locus having k segregating alleles. With samples from a multiallelic gene the frequencies of alleles are no longer independent, and there is a sampling variance–covariance matrix instead of a simple variance. Long's approach simply stacks the

vectors for successive genes to obtain a large covariance matrix. The F statistics come from a partitioning of the covariance matrix using multivariate analysis of variance (MANOVA). Analyses of several sets of data give reasonable results (Long 1986), but the sensitivity of the method to linkage disequilibrium and sampling error remain to be investigated.

Another approach to multiallelic and multilocus data, suggested by Weir and Cockerham (1984), is to use computer bootstrapping techniques, i.e., resampling from the data set over alleles and genes to obtain an empirical distribution of the F statistics using many subsamples of the complete data set. This method has much intuitive appeal: the confidence intervals are empirical, and problems such as linkage disequilibrium do not violate implicit assumptions of the model.

F Population Structure of Yanomama, Gainj, and Kalam Peoples

Many population studies have been made of the Yanomama Indians of the lowlands of tropical South America. The Yanomama live in villages that expand and split, subsisting primarily by slash-and-burn agriculture. The population density (number of individuals per unit area) of the Yanomama is very low, and the large distances that separate their villages, as well as cultural practices, have led to considerable linguistic differentiation (dialect differences) among villages. The genetic data address the genetic structure of the villages, and the consequences of the village structure on the partitioning of genetic variation among them. In one study, the numbers of genotypes of seven different blood groups and serum protein genes were determined in over 500 individuals from 18 different villages. Using the analogy to analysis of variance, Smouse and Long (1988) estimated the F statistics from the sample as follows:

Locus	F_{IS}	F_{IT}	F_{ST}
Rh	−0.0465	−0.0138	0.0312**
Duffy	−0.0034	0.0329	0.0362**
MN	0.0242	0.0841	0.0614**
Alb	0.0234	0.0463	0.0235**
HPT	−0.0646	0.0176	0.0772**
Gc	0.0234	0.0819	0.0599**
Pgm-1	−0.0303	−0.0182	0.0118**
All loci	−0.0157	0.0271	0.0416**

** means that the estimate is significantly different from 0 at the 1% level. The significant F_{ST} values mean that there is significant differentiation among villages.

The F statistics for the pooled set of loci (all loci) was obtained using the method of Long (1986) employing multivariate analysis of variance. Note that there is variability in the F statistics from gene to gene. Much of this is probably due to sampling effects.

The observation that the pooled F_{IS} is negative may be explained by the greater likelihood of females migrating (the migration rate among villages is about 16%). Sexually asymmetric migration results in an outcrossing effect in the offspring, which yields a negative F_{IS}.

The Yanomama may be contrasted with the Gainj and Kalam tribal groups, which live on the northern edge of Papua New Guinea's central highlands. They are as distinct linguistically as the Yanomama dialects, yet they live in an area smaller than a single cluster of Yanomama villages. The resulting population density is about 100 times that of the Yanomama, and the Gainj and Kalam are thought to be in equilibrium with respect to population size. Local groups of the Gainj and Kalam subsist by slash-and-burn agriculture, in groups consisting of 20–200 people. Both the Gainj and the Kalam practice polygamy (which tends to reduce effective population size), but through marriages attempts are made to establish connections with as many people of diverse areas as possible; close inbreeding is thus avoided. Five blood serum proteins were examined in these people (Long et al. 1986), and the F statistics for these groups (considered as one sample) appear below.

Locus	F_{IS}	F_{IT}	F_{ST}
ACP	0.0457	0.0627	0.0177*
ADA	0.0746	0.1013*	0.0289**
GPT	0.0514	0.0920*	0.0428**
HPT	−0.0420	−0.0230	0.0183
GC	0.0529	0.0659	0.0137
All loci	0.0392	0.0628*	0.0225**

The F_{ST} values are again significant, indicating significant genetic separation among groups. F_{IS} in these populations is positive, which has been interpreted as being due to a within-deme Wahlund effect resulting from the marriage customs (Long 1986).

RACE AND GENETIC DISTANCE

To the population biologist, a **race** is a group of individuals in a species showing closer genetic relationships within the group than to members of other such groups. Populations that have undergone some degree of genetic divergence as measured by, for example, F_{ST}, therefore qualify as races. Using this definition, the human population contains many races. Each Yanomama village represents, in a certain sense, a separate "race," and the Yanomama as a whole also form a distinct "race." Such fine distinctions are rarely useful, however; it is usually more convenient to group populations into larger units that still qualify as races in the definition above. These larger units often coincide with races based on physical characteristics such as skin color, hair color and texture, facial features, and body conformation, as defined by

anthropologists. (Modern anthropologists also take cultural and linguistic similarities into account.) Regardless of the criteria, the magnitude of the differences chosen to separate races is essentially arbitrary.

Table 3 gives data indicating that there is as much genetic differentiation among Yanomama villages, as assessed by F_{ST}, as there is among the three "major" human races (Mongoloid, Negroid, Caucasoid). This finding suggests that there is much more genetic variation *within* than *among* human races. However, the conclusion may be misleading. The data in Table 3 are primarily from genes determining allozymes, and allozyme variation may not be representative of total genomic variation. The conclusion of little racial differentiation is certainly not true for genes influencing skin color, hair color, hair texture, and other traits that most people think of in connection with the word "race." However, skin color and other prominent characteristics are used to delineate races precisely because we are aware of these phenotypic attributes, and the genes involved in determining such prominent phenotypic traits cannot be any more representative of the whole genome than allozyme genes are.

In any case, hierarchical grouping of populations can be accomplished using any good measure of the amount of genetic divergence (often called **genetic distance**) between populations. (Most measures of genetic distance

Table 3. Total heterozygosity, average heterozygosity among subpopulations, and fixation index for various organisms.

Organism	Number of populations	Number of loci	Total heterozygosity (H_T)	Subpopulation heterozygosity (H_S)	Fixation index[a] (F_{ST})
Human	3 (major races)	35	0.130	0.121	0.069
Human, Yanomama Indians	37 (villages)	15	0.039	0.036	0.077
House mouse (*Mus musculus*)	4	40	0.097	0.086	0.113
Jumping rodent (*Dipodomys ordii*)	9	18	0.037	0.012	0.676
Drosophila equinoxialis	5	27	0.201	0.179	0.109
Horseshoe crab (*Limulus*)	4	25	0.066	0.061	0.076
Lycopod plant (*Lycopodium lucidulum*)	4	13	0.071	0.051	0.282

[a] F_{ST} is calculated as $(H_T - \bar{H}_S)/H_T$.
(Data from Nei 1975.)

are related to F_{ST} in one way or another, and some of these are discussed in Box G.) Once a suitable measure of genetic distance is chosen, the genetic distance is calculated between all possible pairs of populations. For n populations, there are $n(n-1)/2$ such pairs. Then the two populations with the smallest distance are grouped together and treated as a single "population," and the genetic distance between this new "population" and each of the remaining ones is calculated. The process is repeated until only two such groups remain. The procedures are spelled out for a numerical example in Box G.

Genetic Distance

G

Many measures of genetic distance have been proposed (Edwards 1971; Crow and Denniston 1974; Hedrick 1975; Nei 1977; Smith 1977), but most of them can be related to F_{ST} in some way or another. In this box we will examine Edwards's d and Nei's measure D (not to be confused with D in linkage disequilibrium).

Edwards's d is a measure of genetic divergence between populations based on a geometric model; the reasoning can be illustrated by considering a single gene with two alleles in each of two populations. Let the allele frequencies be denoted p_1 and q_1 in one population, and let the corresponding frequencies in the other population be denoted p_2 and q_2. Geometrically, any two populations fall on the circumference of a quarter-circle of radius 1 with axes $\sqrt{p}$ and $\sqrt{q}$ (Figure 10). The length of a chord connecting the two populations is $\sqrt{2}d$, where $d = \sqrt{[1 - \cos(\theta)]}$, with θ being the angle between the corresponding radii. The quantity d serves as a satisfactory measure of genetic distance, because it can be shown that $d^2 = 1 - \sqrt{(p_1p_2)} - \sqrt{(q_1q_2)}$. Further, if the allele frequencies are not too different, then $2d^2 \approx F_{ST}$.

Perhaps the most widely used measure of genetic distance is the D value of Nei (1975). The basis of the measure is a so-called normalized identity, which expresses the probability that a randomly chosen allele from each of two different populations will be identical, relative to the probability that two randomly chosen alleles from the same population will be identical. ("Identical" in this context means "indistinguishable," not "identical by descent.") The situation is illustrated in Table 4, which shows the alleles of a gene and the allele frequencies that occur in two hypothetical populations, denoted X and Y. Define J_{XY} as the probability that two alleles chosen at random from population X are identical; obviously

$$J_{XX} = \Sigma p_i^2 = 0.455$$

since, with random mating, J_{XX} equals the homozygosity in population X. Likewise, J_{YY} is defined as the probability that two alleles chosen at random from population Y are identical, and in the example

$$J_{YY} = \Sigma q_i^2 = 0.460$$

Now define J_{XY} as the probability that two alleles are identical when one allele is chosen from population X and the other is chosen from population Y. In this example

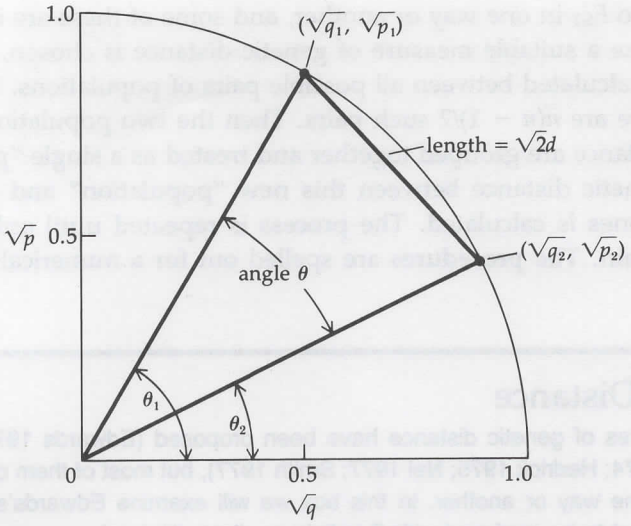


Figure 10. A geometric interpretation of Edwards's genetic distance measure, *d*.

$$J_{XY} = \Sigma p_i q_i = 0.360$$

Nei defines the normalized identity, *I*, for this gene as

$$I = \frac{J_{XY}}{\sqrt{(J_{XX}J_{YY})}} = \frac{0.360}{\sqrt{(0.455)(0.460)}} = 0.7869$$

and the **standard genetic distance**, *D*, as

$$D = -\ln(I) = -\ln(0.7869) = 0.2397$$

Table 4. Frequencies of five alleles at a locus in each of two hypothetical populations.

Alleles	Population X	Population Y
1	$p_1 = 0.60$	$q_1 = 0.30$
2	$p_2 = 0.30$	$q_2 = 0.60$
3	$p_3 = 0.05$	$q_3 = 0$
4	$p_4 = 0.05$	$q_4 = 0$
5	$p_5 = 0$	$q_5 = 0.10$

See the text for calculation of the normalized identity (*I*) and the standard genetic distance (*D*) for this locus.

This measure of genetic distance is more reliable when data from many genes are used. With many genes, the quantities J_{XX}, J_{YY}, and J_{XY} are the arithmetic means of the individual values. Monomorphic loci are included in the analysis. Using the average J's, the same formulas for I and D apply. If two populations are identical, then $J_{XY} = J_{XX} = J_{YY}$, and the normalized identity is 1. Lack of genetic divergence between populations results in a distance $D = 0$. There is no upper bound on D. Since the statistic is calculated from samples from populations, the exact value of D must vary from one sample to another. The degree to which different samples produce different D values depends on the amount of genetic divergence, the magnitude of D, the number of genes, and, of course, the sample sizes. Nei (1978) discusses the sampling properties of D.

One might expect that pairs of populations that are situated farther apart geographically would show greater genetic distance, assuming roughly constant rates of migration. This principle, known as **isolation by distance**, is illustrated in Figure 11, showing the correlation between geographic and genetic distance (Nei's D) based on protein variation in *Drosophila melanogaster* and *D. pseudoobscura* (Singh and Rhomberg 1987b). The observation that greater geographic distances result in less differentiation distance in *D. pseudoobscura* is consistent with other studies that reveal the impressive long-range migration ability of this species.

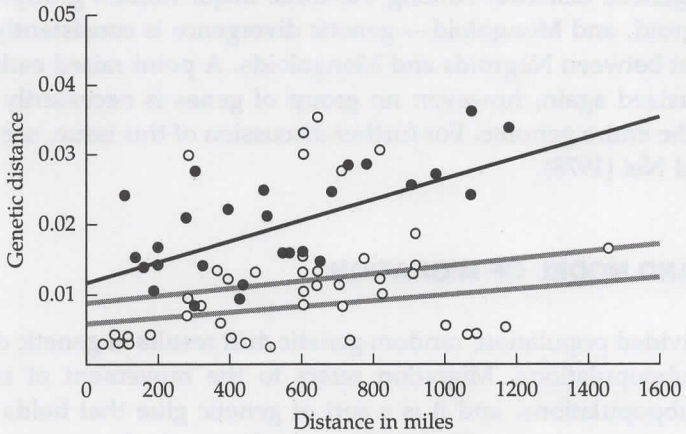


Figure 11. Regression of Nei's genetic distance on geographic distance among populations of *Drosophila melanogaster* (solid dots; black graph) and *D. pseudoobscura* (open circles; gray graphs). The top graph of *D. pseudoobscura* includes all populations; the lower trace excludes the population at Nelson Ranch, Colorado (top seven points). (From Singh and Rhomberg 1987b.)

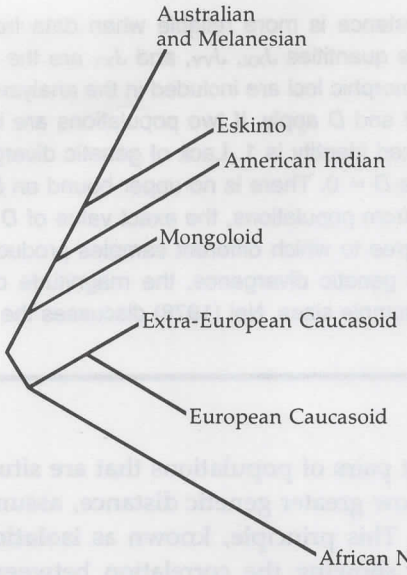


Figure 12. Genetic distance between human populations based on allele frequencies. Each line segment is proportional to genetic distance. (After Cavalli-Sforza 1966.)

Results of a genetic divergence analysis of human allele frequency data are presented in Figure 12, where the length of each line segment is proportional to genetic distance. Among the three major human groups — Caucasoid, Negroid, and Mongoloid — genetic divergence is consistently found to be greatest between Negroids and Mongoloids. A point raised earlier should be emphasized again, however: no group of genes is necessarily representative of the entire genome. For further discussion of this issue, see Lewontin (1972) and Nei (1978).

THE ISLAND MODEL OF MIGRATION

In a subdivided population, random genetic drift results in genetic divergence among subpopulations. Migration refers to the movement of individuals among subpopulations, and it is a sort of genetic glue that holds subpopulations together and sets a limit to how much genetic divergence can occur. To understand the homogenizing effects of migration, it is useful to study migration in several simple models of population structure.

In the island model of migration, a large population is split into many subpopulations dispersed geographically, like islands in an archipelago. Ex-

amples of island population structure might include fish in freshwater lakes or slugs in dispersed garden plots. Each subpopulation is initially assumed to be so large that random genetic drift can be neglected. Consider two alleles A and a with average allele frequencies among the subpopulations equal to $\bar{p}$ and $\bar{q}$, respectively. Migration is assumed to occur in such a way that the migrants are representative of the subpopulations in terms of allele frequency. That is, among the migrants, the allele frequencies of A and a equal $\bar{p}$ and $\bar{q}$, respectively. The amount of migration is measured by a number m which equals the probability that a randomly chosen gene in any subpopulation comes from a migrant. Focus attention on a randomly chosen gene in any subpopulation in generation t. This allele could have come from the same subpopulation in generation $t - 1$ (with probability $1 - m$), in which case it is an A allele with probability p_{t-1}, where p_{t-1} represents the allele frequency of A in the subpopulation in question in generation $t - 1$. Alternatively, the allele could have come from a migrant in generation $t - 1$ (with probability m), in which case it is an A allele with probability $\bar{p}$. (Since random genetic drift, mutation, and natural selection are being ignored, $\bar{p}$ stays the same in all generations.) Altogether,

$$p_t = p_{t-1}(1 - m) + \bar{p}m \tag{6.16}$$

Equation 6.16 is similar to Equation 3.1 for mutation, and its solution in terms of p_0 is

$$p_t = \bar{p} + (p_0 - \bar{p})(1 - m)^t \tag{6.17}$$

where p_0 is the initial frequency of A in the subpopulation in question. As an example of the use of Equation 6.17, suppose there are only two populations with initial allele frequencies of A of 0.2 and 0.8, respectively, with $m = 0.10$ (i.e., 10% of the individuals in either subpopulation in any generation are migrants in which the allele frequency of A is $\bar{p} = (0.2 + 0.8)/2 = 0.5$). What is the allele frequency of A in the two populations after 10 generations? For the population with initial allele frequency 0.2, we substitute $p_0 = 0.2$, $\bar{p} = 0.5$, and $m = 0.10$ into Equation 6.17 to obtain $p_{10} = 0.5 + (0.2 - 0.5)(1 - 0.10)^{10}$, or $p_{10} = 0.395$; for the other population, we substitute $p_0 = 0.8$, $\bar{p} = 0.5$, and $m = 0.10$, so $p_{10} = 0.5 + (0.8 - 0.5)(1 - 0.10)^{10} = 0.605$.

Another example using Equation 6.17 is shown in Figure 13, where there are five subpopulations (initial frequencies 1, 0.75, 0.50, 0.25, and 0), again with $m = 0.10$. Note how rapidly the allele frequencies converge to the same value (in this case, 0.5). Although Equation 6.17 for migration is mathematically similar to Equation 3.1 for mutation, the biological implications are quite different. Because rates of migration are typically much greater than rates of mutation, changes in allele frequency generally occur much faster with migration.

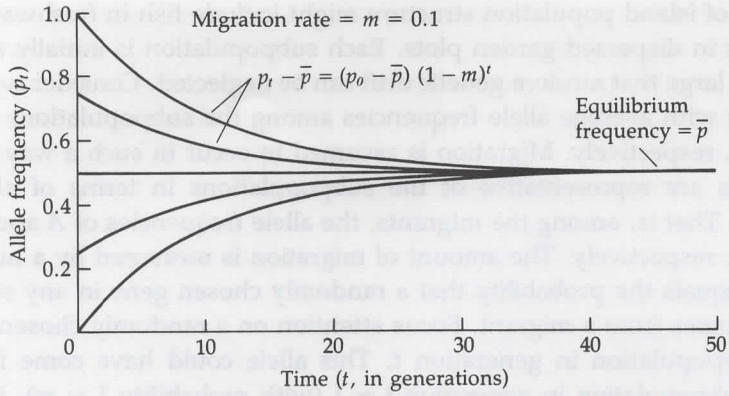


Figure 13. Change of allele frequency against time in five subpopulations exchanging migrants at the rate $m = 0.1$ per generation. Note the rapid convergence to a common equilibrium frequency.

One-Way Migration

When migration occurs predominantly from one population into another without an equal amount of migration in the reverse direction, then $\bar{p}$ in Equation 6.17 should be interpreted as the allele frequency in the population that supplies the migrants. An example is found in racial admixture in the United States. In this case, genetic migration occurs mainly from whites to blacks because individuals of mixed racial ancestry are regarded as blacks. The amount of genetic migration can be estimated using Equation 6.17. The case of the MN blood groups serves as an example. In West Africa (which may, for the purpose of this example, be regarded as the ancestral black population), $\langle p_0 \rangle = 0.474$ for the allele frequency of M. In present-day blacks in Claxton, Georgia, $\langle p_t \rangle = 0.484$. The Claxton white population may reasonably be regarded as representative of the source of the migrants, and for Claxton whites $\langle \bar{p} \rangle = 0.507$. Blacks came to the United States on a large scale from West Africa about 300 years ago, so $\langle t \rangle$ is about 10 generations. Thus, substituting these estimates into Equation 6.17, we obtain $0.484 = 0.507 + (0.474 - 0.507)(1 - \langle m \rangle)^{10}$, from which $1 - \langle m \rangle = [(0.484 - 0.507)/(0.474 - 0.507)]^{1/10} = 0.965$, or $\langle m \rangle = 0.035$ per generation. This estimate can be interpreted as implying that in the genetic history of the population of Claxton blacks, about 3.5% of the alleles of the *MN* gene in any generation were newly introduced by genetic migration from whites. For other genes the estimates of migration rates are: Ss blood groups, $\langle m \rangle = -0.013$ per generation; Duffy blood groups, $\langle m \rangle = 0.011$; Kidd blood groups, $\langle m \rangle = -0.028$; Kell blood groups, $\langle m \rangle = -0.005$; G6PD deficiency, $\langle m \rangle = 0.039$; and for hemoglobin-β, $\langle m \rangle = 0.071$. This example illustrates an important problem

in estimating racial admixture from allele frequencies. The positive values of $\langle m \rangle$ vary widely, and the negative values are not consistent with the proposed model of migration. Cavalli-Sforza and Bodmer (1971) remark that

> The weakness of the analysis is mostly due to the uncertainty of the origin of black Americans . . . and the variability of gene frequencies in the probable area of the slave markets in West Africa. In addition, it is unavoidable that gene frequencies have changed somewhat from their original values, due to drift or, in some cases, selection. The opportunities for admixture, and the time available for it, must also have varied widely.

The most reliable gene among those listed is probably that for the Duffy blood groups, because the Fy^a allele is virtually nonexistent in all of West Africa. For this gene, the estimate of $\langle m \rangle$ is about 1% per generation, which is consistent with the average value for a large number of other genes (Cavalli-Sforza and Bodmer 1971; Wijsman and Cavalli-Sforza 1984).

One-Way Migration and Selection

A model that was studied in some detail by Wright (1931, 1969) considered a continental population that was fixed for one allele, A, and an island population that was polymorphic for alleles A and a. On the island there is natural selection acting on the polymorphism, with fitnesses

AA	Aa	aa
w_{11}	w_{12}	w_{22}

Each generation a fraction m of the population on the island consists of migrants from the mainland, and since the mainland has only the A allele, the effect of migration on the frequency of allele a on the island is

$$\Delta q_{\text{mig}} = -mq \tag{6.18}$$

Using the selection theory from Chapter 4, the change in allele frequency from selection on the island is

$$\Delta q_{\text{sel}} = \frac{pqw_{12} + q^2w_{22}}{p^2w_{11} + 2pqw_{12} + q^2w_{22}} - q \tag{6.19}$$

At equilibrium, the opposing forces of selection and migration must balance each other, so that $\Delta q_{\text{mig}} = -\Delta q_{\text{sel}}$. This relation leads to some messy algebra resulting in a cubic equation in q. One root of the cubic is zero, and the other two are

$$\frac{\pm\sqrt{\{w_{22}^2 - [4m(m+1)w_{11} + 2w_{12}]w_{22} + (2m+1)^2w_{12}^2\}} - w_{22} + (2m+3)w_{12} - 2(m-1)w_{11}}{2(m+1)(w_{11} - 2w_{12} + w_{22})} \tag{6.20}$$

There can be two valid equilibria, but only one can be stable. Two special cases are instructive. The first assumes complete dominance of the A allele with selection favoring the recessive a allele on the island. In this case, the fitnesses are $w_{11} = 1$, $w_{12} = 1$, and $w_{22} = 1 - s$, where $s < 0$. The equilibria, found after substitution into Equation 6.20, are

$$\hat{q} = \frac{s \pm \sqrt{s^2 + 4ms(m + 1)}}{2(m + 1)s} \tag{6.21}$$

An example of the balance between migration and selection against a dominant migrant allele is shown in the top panel of Figure 14. For some values of m, both equilibria in Equation 6.21 are valid, and the graph shows clearly why only one equilibrium is stable.

The second example assumes that the favored allele on the island is dominant. In this case, $s < 0$ and the fitnesses are $w_{11} = 1$, $w_{12} = 1 - s$, and $w_{22} = 1 - s$. Now the equilibria satisfy

$$\hat{q} = \frac{-2s(m + 1) \pm \sqrt{4ms(m + 1)(s - 1)}}{2(m + 1)s} \tag{6.22}$$

The bottom panel of Figure 14 illustrates this case. When the favored allele is dominant, the equilibrium frequency of the migrant recessive allele is increased for any given level of migration. This result makes intuitive sense, because the migrant allele is masked in heterozygotes.

Migration in Finite Populations (Wright's Island Model)

It is remarkable how little migration is required to prevent substantial genetic divergence among subpopulations resulting from random genetic drift. The effect can be seen quantitatively by considering the model of random genetic drift depicted in Figure 2 of Chapter 2, but permitting migration at a rate m according to the island model. The expression for F_t derived in Figure 2 in Chapter 2 is still valid, provided that neither allele α_i or allele α_j has been replaced by a migrant allele. Therefore, in a population of effective size N,

$$F_t = [1/2N + (1 - 1/2N)F_{t-1}](1 - m)^2 \tag{6.23}$$

At equilibrium, $F_t = F_{t-1} = \hat{F}$, so $\hat{F} = [1/2N + (1 - 1/2N)\hat{F}](1 - m)^2$, which, for reasonably small values of m, is approximately

$$\hat{F} = \frac{1}{4Nm + 1} \tag{6.24}$$

Equation 6.24 for migration is plotted in Figure 15. The equation has the same form as Equation 3.10 for mutation, except that Nm replaces $N\mu$. This similarity emphasizes the theoretical relation between the effects of mutation

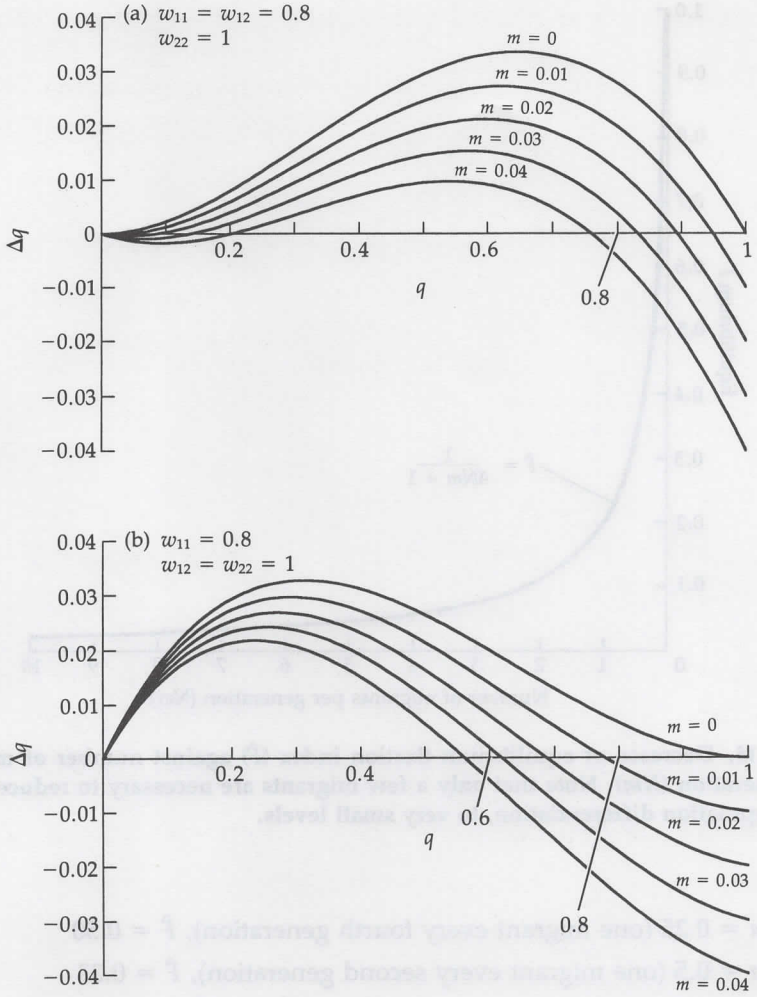


Figure 14. Plots of Δq versus q for populations in migration-selection balance in an island-continent model. In both (a) and (b) allele a is favored by natural selection on the island, and migrants are exclusively of genotype AA. Five levels of migration are considered, from $m = 0$ to $m = 0.04$. In (a) the A allele is dominant, and in (b) the B allele is recessive.

and migration. However, it must be emphasized that rates of migration are typically much greater than rates of mutation, so the practical implications of Equations 3.10 and 6.24 for homozygosity are very different.

Since the actual number of migrants per generation equals Nm, Equation 6.24 implies that $\hat{F}$ decreases as the number of migrants increases. Indeed, the decrease in $\hat{F}$ with increasing Nm is extremely rapid. When $Nm = 0$, of course, $\hat{F} = 1$. For other values of Nm we get

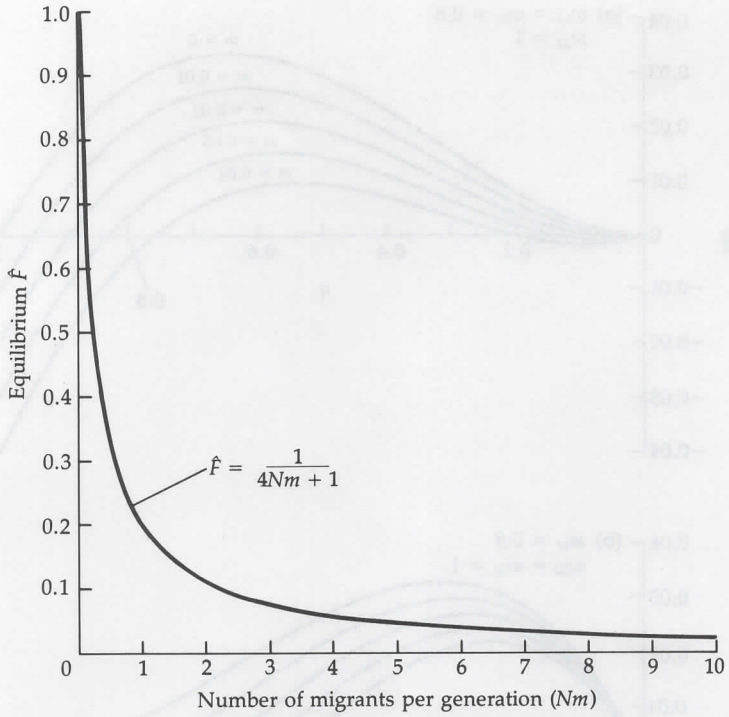


Figure 15. Decrease of equilibrium fixation index ($\hat{F}$) against number of migrants per generation (Nm). Note that only a few migrants are necessary to reduce $\hat{F}$, and thus population differentiation, to very small levels.

$Nm = 0.25$ (one migrant every fourth generation), $\hat{F} = 0.50$

$Nm = 0.5$ (one migrant every second generation), $\hat{F} = 0.33$

$Nm = 1$ (one migrant every generation), $\hat{F} = 0.20$

$Nm = 2$ (two migrants every generation), $\hat{F} = 0.11$

The implication is that migration is a potent force acting against genetic divergence resulting from random genetic drift among subpopulations.

Although the analysis presented above gives an accurate measure of $\hat{F}$ as a function of the migration rate, it is also of interest to determine how migration affects the allele frequency function $\phi(x,t)$, encountered in Chapter 2, which measures the frequency of populations having a particular allele frequency. The diffusion equation for ϕ for an island population that receives migrants at a rate m per generation is

$$\partial\phi/\partial t = \frac{1}{4N_e}\frac{\partial^2}{\partial x^2}[x(1-x)\phi] - m\frac{\partial}{\partial x}[(x_m - x)\phi] \qquad (6.25)$$

assuming no mutation or selection, where the frequency of allele A among migrants is x_m. Note that the second term reduces the rate of change of ϕ by an amount that depends on the migration rate. The greater the migration rate, the more the island is swamped by the migrants, so there is a smaller effect of random drift on the change in allele frequency. The steady state solution to Equation 6.25, discussed at length by Crow and Kimura (1970), is

$$\phi(x) = \left(\frac{\Gamma(4Nm)}{\Gamma(4Nmx_m)\Gamma[4Nm(1 - x_m)]} \right) [x^{4Nmx_m - 1}(1 - x)^{4Nm(1-x_m)-1}] \quad \textbf{(6.26)}$$

The form of ϕ is presented in Figure 16 for a range of values of Nm in the special case where x_m (the allele frequency among the migrants) is 0.5. Note that for small Nm (foreground) random drift predominates, and most of the populations are fixed or nearly fixed. With large Nm, implying either a large population with minimal variance due to random drift, or a high migration rate, there is a tendency for populations to maintain an allele frequency

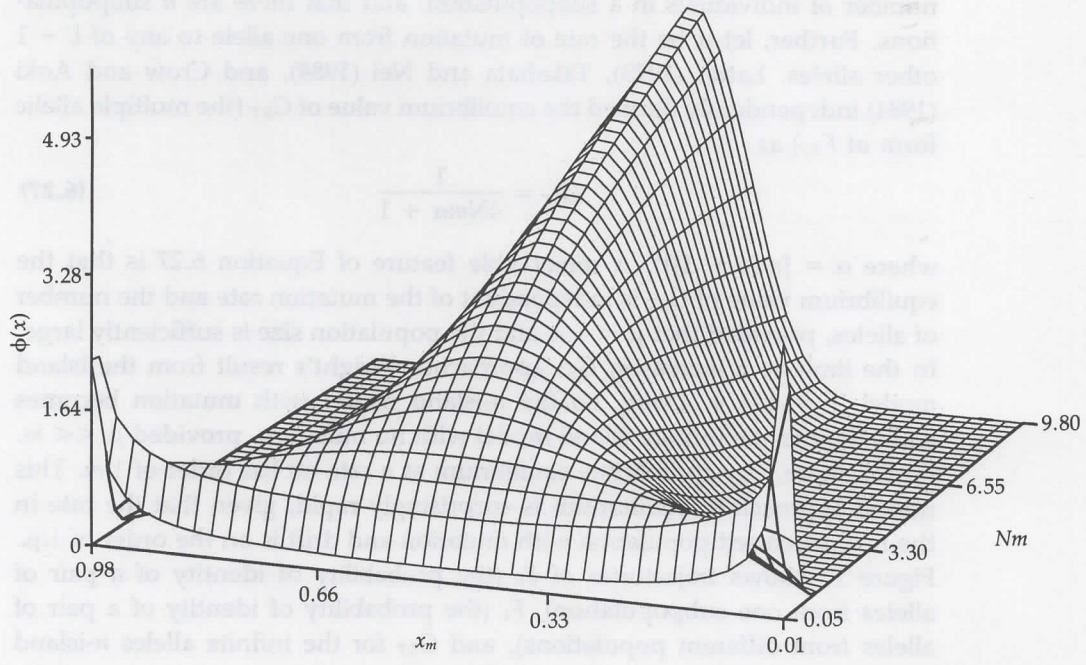


Figure 16. Wright's island model. The solution to the diffusion approximation (Equation 6.26) gives $\phi(x)$, the proportion of populations having allele frequency x, for a range of levels of migration Nm. With very low levels of migration, most populations are near fixation, but with high levels of migration (where migrants have an allele frequency $x_m = 0.5$), the populations tend to have intermediate allele frequencies.

closer to 0.5. Equation 6.26 is actually the solution for a whole class of models where allele frequency is affected by what Wright called *linear pressures*, in which the change in allele frequency is proportional to the allele frequency, as for migration and mutation. It is because they are both linear pressures that the solutions to the island model bear such a strong resemblance to models with mutation.

The Island Model with Mutation

One means of following the consequences of population subdivision is to formulate the recurrence equations for the F statistics in order to determine their changes over time as functions of mutation rate, migration rate, subdivision size, and migration pattern. Much progress was made using this approach by Latter (1973) and Crow and Aoki (1982, 1984), which revealed some important properties of subdivided populations. This approach is also discussed in Crow (1986).

Assume that m is the fraction of genes that are from migrants, N is the number of individuals in a subpopulation, and that there are n subpopulations. Further, let μ be the rate of mutation from one allele to any of $k - 1$ other alleles. Latter (1973), Takahata and Nei (1984), and Crow and Aoki (1984) independently derived the equilibrium value of G_{ST} (the multiple allelic form of F_{ST}) as

$$\hat{G}_{ST} = \frac{1}{4Nm\alpha + 1} \tag{6.27}$$

where $\alpha = [n/(n - 1)]^2$. A remarkable feature of Equation 6.27 is that the equilibrium value of G_{ST} is independent of the mutation rate and the number of alleles, provided that $\mu << m$ and the population size is sufficiently large. In the limit as n increases, G_{ST} approaches Wright's result from the island model [$F = 1/(1 + 4Nm)$], so the n-island model with mutation becomes equivalent to the infinite island model with no mutation, provided $\mu << m$. Furthermore, G_{ST} approaches equilibrium at a rate on the order of $1/m$. This rate of approach to equilibrium is surprisingly rapid, given that the rate in the nonstructured population with mutation and drift is on the order of $1/\mu$. Figure 17 shows trajectories of F_0 (the probability of identity of a pair of alleles from one subpopulation), F_1 (the probability of identity of a pair of alleles from different populations), and G_{ST} for the infinite alleles n-island model. The relevant equations are (Crow 1986, p. 64)

$$\begin{aligned} F_0' &= \{a[c + (1 - c)F_0] + (1 - a)F_1\}(1 - \mu)^2 \\ F_1' &= \{b[c + (1 - c)F_0] + (1 - b)F_1\}(1 - \mu)^2 \\ F &= [F_0 + (n - 1)F_1]/n \\ G_{ST} &= (F_0 - F)/(1 - F) \end{aligned} \tag{6.28}$$

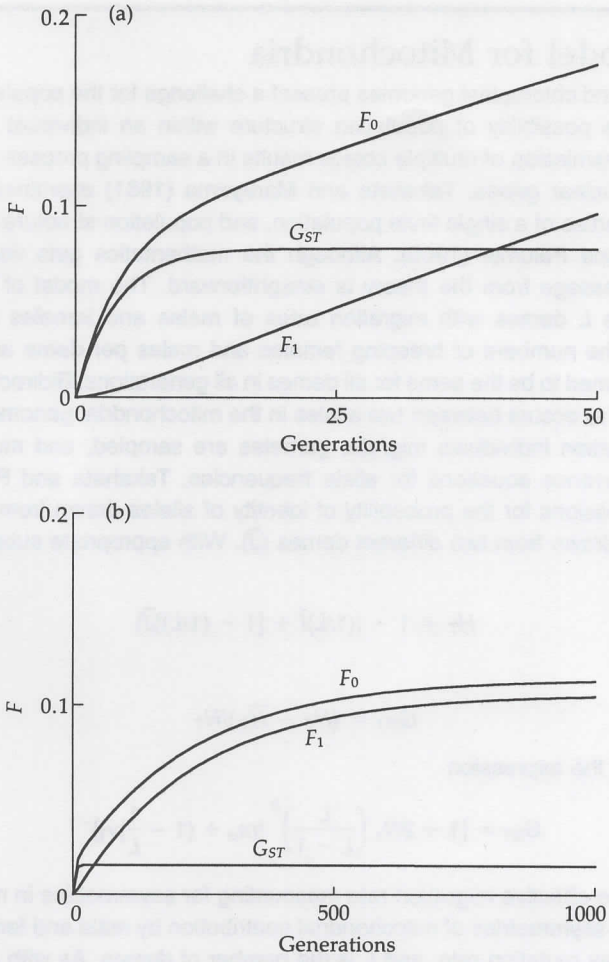


Figure 17. The changes in F_0, F_1, and G_{ST} in the n-island model according to Equations 6.28. For both (a) and (b) the number of individuals per island was $N = 20$, the number of islands was $n = 10$, the mutation rate was $\mu = 0.001$, and the migration rate was 0.1. (a) The first 50 generations of the simulation, starting with all identity measures equal to zero, and demonstrating how rapidly G_{ST} attains equilibrium. (b) The figure verifies that the other identity measures also attain a steady-state, but after a much longer time.

where $a = (1 - m)^2 + m^2/(n - 1)$, $b = (1 - a)/(n - 1)$, and $c = 1/2N_e$. Figure 17 demonstrates that G_{ST} approaches equilibrium faster than the other measure of inbreeding. Crow and Aoki (1984) emphasize that the rapid attainment of equilibrium and the relative insensitivity to mutation rate make G_{ST} a useful measure for population genetic analysis.

H Island Model for Mitochondria

Mitochondrial and chloroplast genomes present a challenge for the population geneticist because of the possibility of population structure within an individual cell; also, the uniparental transmission of multiple copies results in a sampling process much different from that of nuclear genes. Takahata and Maruyama (1981) examined many of the sampling properties of a single finite population, and population structure was dealt with by Takahata and Palumbi (1985). Although the mathematics gets very heavy very quickly, the message from the theory is straightforward. The model of Takahata and Palumbi allows L demes with migration rates of males and females of m and m^*, respectively. The numbers of breeding females and males per deme are N_f and N_m, which are assumed to be the same for all demes in all generations. Bidirectional mutation at rates v_1 and v_2 occurs between two alleles in the mitochondrial genome. During each discrete generation individuals migrate, gametes are sampled, and mutation occurs. From the recurrence equations for allele frequencies, Takahata and Palumbi (1985) obtained expressions for the probability of identity of alleles drawn from one deme ($\bar{I}$), and of alleles drawn from two different demes ($\bar{J}$). With appropriate substitutions using $\overline{H}_S = 1 - \bar{I}$,

$$H_T = 1 - \{(1/L)\bar{I} + [1 - (1/L)]\bar{J}\}$$

and

$$G_{ST} = (H_T - \overline{H}_S)/H_T$$

they arrived at the expression

$$G_{ST} = \{1 + 2N_e\left(\frac{L}{L-1}\right)^2 [m_e + (1 - \frac{1}{L})v]\}^{-1}$$

where m_e is the effective migration rate (accounting for asymmetries in migration of the two sexes and asymmetries of mitochondrial contribution by male and female gametes), v is the one-way mutation rate, and L is the number of demes. As with nuclear genes, populations become more differentiated with smaller deme size, lower migration rate, and lower mutation rate. An interesting difference from nuclear genes is that the relative contribution of mutation and migration to population differentiation depends on the number of demes.

All of the models of subdivided populations that have been discussed consider the population structure to be fixed in that the number of demes and the rates of migration are assumed to be constant. A natural alternative formulation considers extinction and recolonization of local subpopulations. The process of founding new subpopulations from migrants can result in an important source of gene flow and, under appropriate conditions on the parameters, results analogous to Equation 6.27 have been obtained (Slatkin 1977; Maruyama and Kimura 1980).

STEPPING STONE MODELS

When populations are organized in discrete colonies or demes with migration between them, there is a discrete structure to the population that is distinct from that considered in Wright's isolation by distance models. Such inter-deme migration models are called *stepping stone models*; it is often assumed that only migration into adjacent demes is allowed. The geometrical pattern of demes may exist in from 1 to n dimensions, and it may be linear, circular, rectangular, spherical, torroidal, and so on. Circular models avoid the problems of edge effects caused by the unidirectional migration of individuals in demes at the ends of linear and planar patterns.

Kimura and Weiss (1964) considered stepping stone models with a linear array of demes and with two- and three-dimensional lattices of demes (see Figure 18). They allowed migration to occur only between adjacent demes, and calculated the correlation coefficient of alleles sampled from a distance d apart as

$$\begin{aligned} &\text{1 dimension} && r(d) \propto \exp[-\sqrt{(2m_\infty/m_1)}d] \\ &\text{2 dimensions} && r(d) \propto (1/\sqrt{d})\exp[-\sqrt{(4m_\infty/m_1)}d] \\ &\text{3 dimensions} && r(d) \propto (1/\pi d)\exp[-\sqrt{(6m_\infty/m_1)}d] \end{aligned} \qquad \textbf{(6.29)}$$

where m_∞ is a long-distance migration parameter referring to the probability that a migrant comes from the population at large. Long distance migration is formally equivalent to mutation, as in the island model. The parameter m_1 is the probability of migration from an adjacent island. The significant point in Equations 6.29 is that the correlation falls off much more rapidly as additional dimensions are added to a stepping stone model (Figure 19). This suggests that populations in linear habitats, such as along shorelines, might be expected to undergo greater genetic differentiation per unit length than those in two-dimensional habitats, assuming equal rates and distances of migration. Slatkin (1985b) expressed the results of Kimura and Weiss (1964) in terms of the ratio of the probability that a pair of alleles from different subpopulations are identical by descent (F_1) compared to that of alleles within a subpopulation (F_0). In the case of one dimension, with $\mu << m$, he obtained $F_1/F_0 = 1 - \sqrt{2\mu/m}$. For two dimensions the ratio is $1 - \pi\ln[4\sqrt{(\mu/m)}]/2$. In both cases the critical parameter determining isolation by distance is the ratio μ/m.

In a series of papers, Maruyama (1969, 1970, 1971, 1972b,c) discovered several properties of stepping stone models. One observation was that the average number of heterozygotes bearing a particular allele that occur in a population before the loss or fixation of the allele is $2N$, where N is the population size. This is true regardless of population structure. Similarly, the

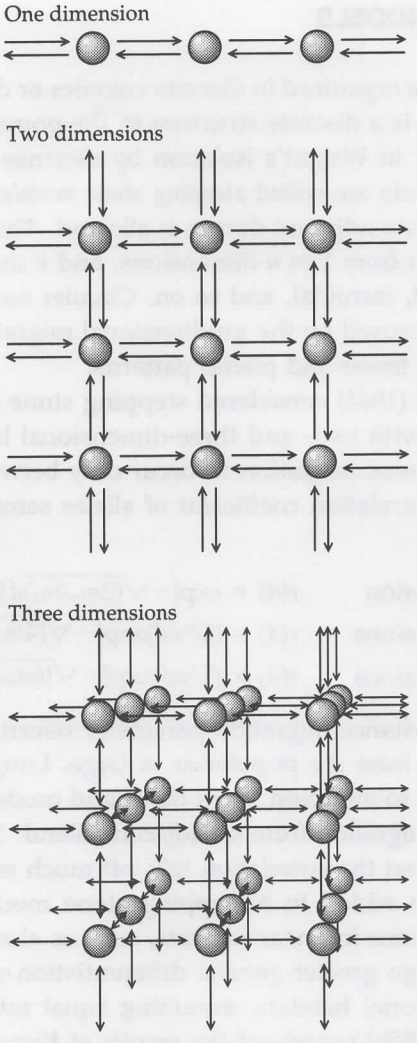


Figure 18. Patterns of migration in the one-, two-, and three-dimensional stepping stone model. Each arrow represents a migration rate of $m/2$ in one dimension, $m/4$ in two dimensions, and $m/6$ in *each direction* in three dimensions. Edge effects can be eliminated by wrapping the one-dimensional array into a circle, the two-dimensional array into the surface of a sphere, and the three-dimensional array into a torus.

fixation probability of a mutation is unaffected by population subdivision. Kimura and Maruyama (1971) demonstrated an important similarity between two-dimensional stepping stone models and the island model — that populations will be strongly differentiated if $Nm << 1$; whereas if $Nm > 4$, they behave as a single panmictic unit.

For the most part, stepping stone models yield very complex mathematics

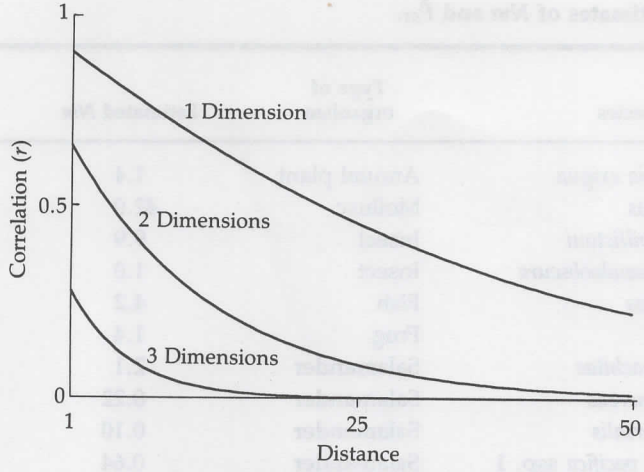


Figure 19. The decline in the correlation of allele frequencies under the one-, two-, and three-dimensional stepping stone model. (After Kimura and Weiss 1964.)

and many of the results are not easy to apply to field or experimental populations. On the other hand, some results are remarkably simple, and the predictions are commonly observed by population geneticists working with samples from natural populations. The best known result of this kind is the observation of clines of allele frequency along a geographic transect. Clines can be formed by a variety of processes, including differential selection at the two ends (combined with migration of the isolation by distance sort, or by migration with the allele frequency differences at the ends due to historical accident). Some well-studied clines (discussed further in Chapter 9) include the cline in *Adh* frequencies in *Drosophila melanogaster*, the *Ldh* cline in *Fundulus*, the LAP cline in *Mytilus*, and the HLA cline in humans.

ESTIMATING MIGRATION RATES

One method of estimating genetic migration in natural populations relies on the finding that, in theoretical models, the logarithm of Nm decreases approximately as a linear function of the average frequency of private alleles in samples from the subpopulations (Slatkin 1985a,b). *Private alleles* are those found in only one subpopulation. Data on the average frequency of private alleles in samples have been compiled and analyzed by Slatkin (1985b), and the resulting estimates of Nm and equilibrium values of F_{ST} are summarized in Table 5. There is obviously considerable variation in Nm among organisms.

Table 5. Estimates of Nm and $\hat{F}_{ST}$.

Species	Type of organism	Estimated Nm	Estimated $\hat{F}_{ST}$
Stephanomeria exigua	Annual plant	1.4	0.152
Mytilus edulis	Mollusc	42.0	0.006
Drosophila willistoni	Insect	9.9	0.025
Drosophila pseudoobscura	Insect	1.0	0.200
Chanos chanos	Fish	4.2	0.056
Hyla regilla	Frog	1.4	0.152
Plethodon ouachitae	Salamander	2.1	0.106
Plethodon cinereus	Salamander	0.22	0.532
Plethodon dorsalis	Salamander	0.10	0.714
Batrachoseps pacifica ssp. 1	Salamander	0.64	0.281
Batrachoseps pacifica ssp. 2	Salamander	0.20	0.556
Batrachoseps campi	Salamander	0.16	0.610
Lacerta melisellensis	Lizard	1.9	0.116
Peromyscus californicus	Mouse	2.2	0.102
Peromyscus polionotus	Mouse	0.31	0.446
Thomomys bottae	Gopher	0.86	0.225

(Data from Slatkin 1985a.)

However, many of the values of Nm are approximately 2 or smaller, which gives considerable scope for genetic divergence resulting from random genetic drift.

A second approach to estimating Nm in natural populations is illustrated in Figure 20, which gives the distribution of estimated values of $\langle H_S \rangle$, $\langle H_T \rangle$ and $\langle F_{ST} \rangle$ among 61 genes in natural populations of *Drosophila melanogaster* (Singh and Rhomberg 1987b). The average of the estimated values is $F_{ST} = 0.16$, which, assuming equilibrium, estimates $1/(4Nm + 1)$ (Equation 6.24). Therefore, $\langle Nm \rangle = [(1/0.16) - 1]/4 = 1.3$. This estimate is within the range for other *Drosophila* species in Table 5. However, there are many genes in Figure 20 that have $\langle F_{ST} \rangle$ values greater than 0.3.

Rare alleles can be particularly useful in inferences about population structure. We have seen that certain properties of neutral alleles are unaffected by population subdivision, including the expected time to fixation. However, deleterious alleles have a shorter expected time until loss, and many rare alleles are likely to be deleterious. Since rare alleles are more likely to have arisen recently, they are more likely to reveal the effects of population subdivision. This argument has been supported both by theory and by empirical observations of population samples (Kimura 1983).

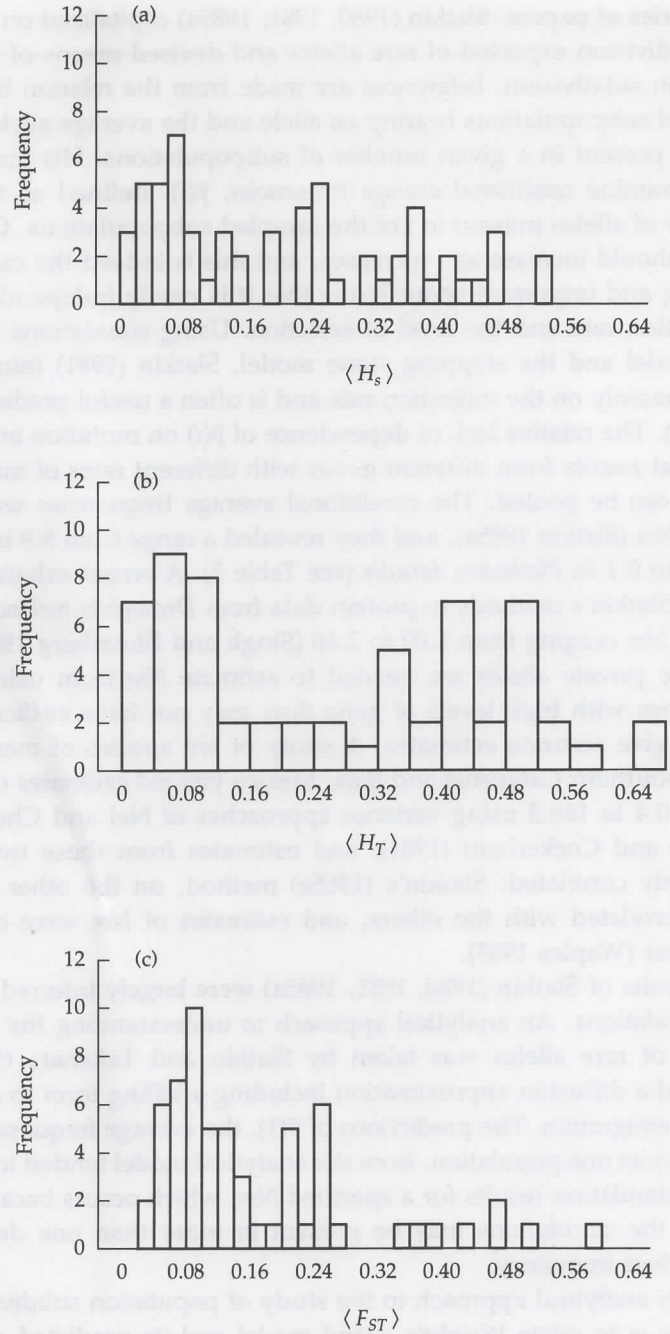


Figure 20. Distribution of (a) single locus heterozygosity (H_S), (b) total genic diversity (H_T), and (c) fixation index (F_{ST}) at polymorphic loci in geographic populations of *Drosophila melanogaster*. (From Singh and Rhomberg 1987b.)

In a series of papers, Slatkin (1980, 1981, 1985a) capitalized on the heightened subdivision expected of rare alleles and devised means of quantifying population subdivision. Inferences are made from the relation between the number of subpopulations bearing an allele and the average allele frequency of alleles present in a given number of subpopulations. His first approach was to examine *conditional average frequencies*, $\overline{p}(i)$, defined as the average frequency of alleles present in i of the sampled subpopulations. One expects that $\overline{p}(i)$ should increase as i increases, and this is indeed the case. What is surprising and important about $\overline{p}(i)$ is that it is nearly independent of both the mutation rate and the level of selection. Using simulations of both the island model and the stepping stone model, Slatkin (1981) found that $\overline{p}(i)$ depends mainly on the migration rate and is often a useful predictor of gene flow (Nm). The relative lack of dependence of $\overline{p}(i)$ on mutation and selection means that results from different genes with different rates of mutation and selection can be pooled. The conditional average frequencies were used to estimate Nm (Slatkin 1985a), and they revealed a range from 9.9 in *Drosophila willistoni* to 0.1 in *Plethodon dorsalis* (see Table 5). A recent exhaustive application of Slatkin's methods to protein data from *Drosophila melanogaster* gave values of Nm ranging from 1.09 to 2.66 (Singh and Rhomberg 1987a).

Because private alleles are needed to estimate Nm from using Slatkin's $\overline{p}(1)$, species with high levels of gene flow may not have sufficient private alleles to give accurate estimates. A study of ten species of marine coastal fishes of southern California and Baja, Mexico yielded estimates of Nm ranging from 0.4 to 166.5 using variance approaches of Nei and Chesser (1983) and Weir and Cockerham (1984), and estimates from these two methods were highly correlated. Slatkin's (1985a) method, on the other hand, was weakly correlated with the others, and estimates of Nm were consistently much lower (Waples 1987).

The results of Slatkin (1980, 1981, 1985a) were largely inferred from computer simulations. An analytical approach to understanding the population structure of rare alleles was taken by Slatkin and Takahata (1985), who considered a diffusion approximation including a killing term to account for loss from emigration. The predictions of $\overline{p}(1)$, the average frequency of alleles present in just one population, from the analytical model tended to be smaller than the simulation results for a specified Nm, which occurs because private alleles in the simulations may be present in more than one deme, hence inflating their frequency.

Another analytical approach to the study of population subdivision using rare alleles is to relate Wright's island model and its predicted distribution $\phi(x)$ to the expected conditional average frequencies. Barton and Slatkin (1986) make the connection by showing that the expected frequency of an allele occurring in a deme (given that it is present) is

$$p = \frac{\int_0^1 x\phi(x)\mathrm{d}x}{(1 - a_o)} \qquad (6.30)$$

where a_o is the probability that the allele is absent from a sample. In deriving Equation 6.30 there is an implicit assumption that the population is in a steady state, or a quasiequilibrium, with respect to the gain and loss of alleles. Tests of this assumption using numerical simulation show that the conditional average frequency attains steady state approximately as rapidly as F_{ST}—which means on the order of $1/m$ generations.

INTERDEME SELECTION AND THE SHIFTING BALANCE THEORY

One type of selection should be singled out for brief discussion in the context of migration because it involves an extended concept of "fitness." The type of selection is called **interdeme selection**, and it occurs between semi-isolated populations (demes) of the same species. If populations containing certain genotypes are more likely to become extinct and have their vacated habitats recolonized by migrants from other populations that are more persistent due to the particular genotypes that they contain, then the more successful populations can in some sense be considered as having a greater "fitness" than the less successful ones. Since this concept of population fitness is a characteristic of the entire population and not merely the average fitness of the genotypes within it ($\overline{w}$), interdeme selection is outside the realm of most conventional models of selection. Interdeme selection is one type of group selection (Wilson 1983).

Interdeme selection plays an essential role in the shifting balance theory of evolution (Wright 1977, and earlier). In Wright's view, subdivision of a population into small, semi-isolated demes gives the best chance for the populations to explore the full range of their adaptive topography. Temporary reductions in fitness that would be prevented by selection in large populations become possible in small ones because of the random drift in allele frequencies that occurs in small populations. The lucky subpopulations that reach higher adaptive peaks on the fitness surface increase in size and send out more migrants than other subpopulations, and the favorable gene combinations are gradually spread throughout the entire set of subpopulations by means of interdeme selection. The shifting balance process includes three distinct phases:

1. An *exploratory phase*, in which random genetic drift plays an important role in allowing small populations to explore their adaptive topography.
2. A phase of *mass selection*, in which favorable gene combinations created

by chance in phase 1 rapidly become incorporated into the genome of local populations by the action of natural selection.

3. A phase of *interdeme selection*, in which the more successful demes increase in size and rate of migration, and the excess migration shifts the allele frequencies of nearby populations until they also come under the control of the higher fitness peak. The favorable genotypes thereby become spread throughout the entire population in ever-widening concentric circles. Where the region of spread from two such centers overlaps, a new and still more favorable genotype may occur and itself become a center for interdeme selection. In this manner, the whole of the adaptive topography can be explored, and there is a continual shifting of control from one adaptive peak to a superior one.

The shifting balance theory has played an important role in evolutionary thinking, in part because of the prominent role assigned to random genetic drift in the initial phase of the process. However, as a comprehensive theory of evolution, many aspects of the theory remain to be tested. For the theory to work as envisaged, the interactions between alleles must often result in complex adaptive topographies with many peaks and valleys. The population must be split up into smaller demes, which must be small enough for random genetic drift to be important but large enough for mass selection to fix favorable combinations of alleles. While migration between demes is necessary, neighboring demes must be sufficiently isolated for genetic differentiation to occur, but sufficiently connected for favorable gene combinations to spread. Because of uncertainty about the applicability of these assumptions, the shifting balance process remains a picturesque metaphor that is still largely untested. Discrepancies between predictions of the shifting balance theory and observations of molecular sequence variation were emphasized by Nei (1987, pp. 419–422).

I Human Isolates and the Shifting Balance Theory

The shifting balance theory holds a peculiar place among scientific theories in that many of its components can be explored in some detail, but as a whole it is almost untestable. One prediction of the theory is that selection can lead different isolated populations to different genetic destinations as a result of chance or historical events. An intriguing, and rather speculative, example of this appeared in a study by Zlotogora et al. (1988). Four Israeli communities characterized by large families and high levels of consanguinity were found to display four different lysosomal storage disorders. The coincidence of four distinct diseases with similar biochemical consequences is quite extraordinary. Among the Ashkenazi Jews, four of the seven most common genetic disorders are lysosomal storage disorders. The four disorders — including Tay Sachs, Gaucher, Niemann-Pick, and mucolipidosis IV — are genetically distinct, but all four are

characterized by an excess of a type of lipid known as a sphingolipid, and all involve sphingolipids with a ceramide-lipid backbone. (Sphingolipids are important in the membranes of cells of the central nervous system.) Although it is not useful to perform a formal statistical test on this observation, it does suggest that the clustering of such similar diseases would not be likely to occur by chance alone. There has been a suggestion that the prevalence of lysosomal storage disease in these populations may be the result of natural selection, perhaps favoring carriers of the diseases in some as yet undisclosed manner. One speculation is that resistance to tuberculosis and pneumonia is involved. However, the cause of the clustering of these otherwise rare diseases remains an intriguing mystery.

One implication of interdeme selection is that selection for genes that are beneficial to the group may, in some cases, override effects that are harmful to the individuals. This principle is illustrated in the model in Table 6, where the allele A' is harmful to individuals within demes but is favorable to the deme as a whole. Equation 4.4 implies that, within the ith deme, $\Delta q_i = -cq_i(1 - q_i)$ (assuming that $\overline{w} \approx 1$). Averaging across all n subpopulations, the change in allele frequency resulting from selection within subpopulations Δq_w equals $-c\bar{q}(1 - \bar{q})(1 - F)$. At the same time within-population selection occurs, interdeme selection favors demes containing A', and the change in allele frequency in the ith population resulting from between-population selection equals $2(b - c)q_i(1 - q_i)$ (assuming $\overline{v} \approx 1$). Averaging this across all n subpopulations gives the total change in allele frequency resulting from between-population selection as $\Delta q_b = 2(b - c)\bar{q}(1 - \bar{q})(1 - F)$. Putting both processes together, the total change in the frequency of A' is

$$\Delta q = \Delta q_w + \Delta q_b = \bar{q}(1 - \bar{q})[2F(b - c) - c(1 - F)] \qquad \textbf{(6.31)}$$

Equation 6.31 implies that $\Delta q > 0$ if

$$\frac{b - c}{c} > \frac{1 - F}{2F} \qquad \textbf{(6.32)}$$

Table 6. Model of interdeme selection.

Genotype	AA	AA'	$A'A'$
Frequency in deme i	p_i^2	$2p_iq_i$	q_i^2
Within-population fitness (w)	1	$1 - c$	$1 - 2c$
Between-population fitness of deme i (v)		$1 + 2(b - c)q_i$	

This is the condition necessary in order for selection between demes to override selection within demes, and the formulation is quite general (Crow and Aoki 1982). (A biological interpretation of Equation 6.32 is given in Chapter 9 in the section on kin selection.)

If there are n demes, each of size N, which exchange migrants in such a way that m is the proportion of genes in each deme that are exchanged each generation for genes chosen at random from the other demes, then the approximate value of G_{ST} at equilibrium was shown in Equation 6.27 to be

$$\hat{G}_{ST} = \frac{1}{4Nm\alpha + 1} \tag{6.33}$$

where $\alpha = [n/(n-1)]^2$ (Crow and Aoki 1984). Note that Equation 6.33 does not depend on the mutation rate to new alleles or on the number of alleles. When n is reasonably large, $\alpha \approx 1$, and Equation 6.33 reduces to the familiar form in Equation 6.24.

Equation 6.33 implies that, for large n, $(1 - F)/2F = 2Nm$, which equals two times the number of migrant diploid individuals per generation. We therefore conclude from Equation 6.32 that selection between demes overrides selection within demes only when the benefit to the group $(b - c)$ relative to the cost to the individual (c) is greater than twice the average number of migrant individuals per generation. This principle defines a rather stringent limit above which migration among demes cancels any possible effects of interdeme selection.

PROBLEMS

1 What value of m in the island model is equivalent to population fusion?

2 If a mainland population of snails has an allele frequency of 0.8 and an island population has a frequency of 0.2 for the same allele, how many generations are required for the island population to achieve an allele frequency of 0.5, given a migration rate of 0.01?

3 If four populations with allele frequencies 0.2, 0.4, 0.6, and 0.8 undergo migration according to the island model with $m = 0.05$, what are the expected allele frequencies after 10 generations?

4 Two diploid random mating populations have allele frequencies $q + \epsilon$ and $q - \epsilon$ for a recessive allele of a gene. What are the frequencies of homozygotes before and after population fusion?

5 What is the analogue of Equation 6.1 for males affected by an X-linked recessive disorder?

6 Show that $F_{IT} = F_{IS} + F_{IT} - F_{IS}F_{IT}$ and interpret the expression.

7 In the island model of migration, how does the variance in allele frequency among populations change as a function of m and t?

8 Suppose two subpopulations with equal allele frequencies of two linked genes have an amount of linkage disequilibrium that is equal but opposite in sign. What is the amount of linkage disequilibrium in a population formed by mixing equal numbers of individuals from the two populations?

9 Calculate F_{ST} among the three random-mating populations below based on the specified allele frequencies. What is the maximum value of F_{ST} in this situation?

		Population		
		1	2	3
Allele	1	0.1	0.2	0.3
	2	0.3	0.3	0.3
	3	0.6	0.5	0.4

10 Considering Equation 6.3, what is the rate at which random genetic drift increases the variance in allele frequency among subpopulations?

11 When random genetic drift is offset by migration among populations in the island model, what value of m is necessary to keep the equilibrium value of F smaller than 0.05?

12 Calculate F_{IS}, F_{ST}, and F_{IT} for the populations with the genotype frequencies shown below.

		Population 1	Population 2
Genotype	*AA*	0.056	0.072
	Aa	0.288	0.256
	aa	0.656	0.672

13 In a population of monoecious plants in Hardy-Weinberg proportions for two alleles with allele frequency p, what is the variance in allele frequency among plants? What is the variance if the population were completely inbred?

14 For a population of plants undergoing self-fertilization, what is the variance in allele frequency among individuals in generation n, assuming $F_0 = 0$. (Hint: See Problem 13.)

15 Consider a gene that is a recessive lethal in an island environment, which is introduced into the island population each generation by migration from a mainland population at the rate m, where m is small. What is the equilibrium frequency of the lethal allele on the island?

16 What are the values of F_{ST}, F_{IS}, and F_{IT} among isolated populations of effective size 50 that undergo

a. 47 generations of random mating followed by 3 generations of sib mating.

b. 50 generations of random mating.

17 In the island model of migration, if m_i is the rate of migration in generation i, what is the analogous equation to Equation 6.17? If Equation 6.17 is applied, how can m be interpreted?

18 The measure of genetic divergence G_{ST} is very useful for multiple alleles in multiple subpopulations. G_{ST} can be defined as $(J_S - J_T)/(1 - J_T)$, where p_i is the frequency of the ith allele, $J_S = \Sigma\text{Avg}(p_i^2)$ and $J_T = \Sigma[\text{Avg}(p_i)]^2$ (Nei 1987). The summation means summation over all alleles, and Avg means the average over all subpopulations. For the random mating populations below, calculate F_{ST} and G_{ST}.

	Population 1	Population 2
1	0.2	0.6
Allele 2	0.3	0
3	0.5	0.4

19 G_{ST} for multiple alleles is actually a weighted average of F_{ST} values,

$$G_{ST} = \frac{\Sigma \bar{p}_i(1 - \bar{p}_i)F_{ST(i)}}{\Sigma \bar{p}_i(1 - \bar{p}_i)}$$

where the summation is over all alleles, $\bar{p}_i$ is the average frequency of the ith allele among the subpopulations, and $F_{ST(i)}$ is the F_{ST} value for the ith allele calculated as if the gene had only two alleles with frequencies p_i and $1 - p_i$ in each subpopulation. Calculate $F_{ST(i)}$ for each allele in Problem 18 and confirm numerically that the weighted average equals G_{ST}.

20 For the populations in Problem 18, calculate the normalized identity I and the standard genetic distance D.

Chapter Seven

Molecular Population Genetics

Evolution is both process and result. The process of evolution includes all mechanisms of genetic change that occur in organisms through time, with special emphasis on those mechanisms that promote the adaptation of organisms to their environment or that lead to the formation of new, reproductively isolated species. The result of evolution is observed in the evolutionary history of life on Earth and in the genetic relationships among species that exist today.

Corresponding to evolution as process and result are two principal themes that recur in evolutionary studies. The first is that of phylogenetic inference, which aims to determine the evolutionary history and genetic relatedness among groups of organisms. Phylogenetic inference emphasizes the comparative side of evolutionary studies. Traditionally it includes comparative morphology, anatomy, and embryology, but modern systematic biologists also emphasize biochemistry, biogeography, ecology, behavior, and other approaches. The second theme of evolutionary studies is that of mechanism, which is primarily within the purview of population genetics. Mechanistic studies emphasize the interplay between population structure, mating systems, migration, mutation, random genetic drift, and natural selection in the microevolutionary processes that lead to genetic change and adaptation.

Techniques used in molecular biology, particularly those for determining amino acid or nucleotide sequences, have provided data that has added a new dimension to phylogenetic inference. For example, studies of differences in DNA sequence among bird species, determined by measuring the thermal stability of hybrid DNA molecules containing one DNA strand from each species, have led to several major revisions in classical bird taxonomy. One unexpected finding from these DNA hybridization studies was that the tribe Corvini, which includes crows, ravens, jays, magpies, and their relatives, originally evolved as part of a larger group in Australia that subsequently

spread to Asia and thence throughout the world (Sibley and Ahlquist 1986). On an even grander scale, the analysis of 5S RNA sequences in a broad variety of microorganisms has led to a reclassification at the deepest of phylogenetic levels, resulting in a new kingdom, the Archaebacteria (Woese 1981).

The great power and relative ease of comparative molecular biology have given prominence to phylogenetic inference. Phylogenetic studies have not, however, eclipsed the search for evolutionary mechanisms. Indeed, the techniques of molecular biology have invigorated the study of evolutionary mechanisms through the discovery of unprecedented types of genes and DNA sequences, the evolutionary implications of which are still to be worked out. At the same time, the new molecular data provide unique opportunities for inferring mechanisms of evolutionary change and testing evolutionary hypotheses. Evolutionary biology is rapidly reaching the point where the understanding of evolution includes not only the litany of which species begat which others, but also an appreciation of how and why such changes came to pass.

This chapter focuses on the molecular aspects of population genetics. It includes discussions of the neutral theory of molecular evolution, the evolution of multigene families, and the evolutionary implications of transposable elements. The chapter demonstrates the important role of molecular biology in recent evolutionary studies: many of the discussions could not have been considered ten years ago because the phenomena at issue had not yet been discovered.

THE TOOLS

Population genetics, like all sciences, is directly affected by the availability of instruments and techniques to obtain the data relevant to the questions being asked. Indeed, to a certain extent, the tools and techniques dictate the questions. This may seem to trivialize the way that science progresses, but in actual practice the frenzy for finding the newest and cleverest ways to solve a given experimental problem greatly accelerates the rate of scientific progress. The tools that are available to the molecular population geneticist today are impressive indeed — particularly compared to the situation only ten years ago.

The first molecular technique applied to questions about the evolution of molecular sequences was the Sanger procedure for determining the amino acid sequence of a protein. The original technique, applied to bovine insulin (1949–1955), was extremely laborious. Today it is far simpler to sequence the DNA of the relevant gene and infer the amino acid sequence of the protein

product. Various electrophoretic procedures for separating proteins on the basis of charge or molecular weight are still very useful and widely used, but because they do not yield information about molecular sequences, they will not be considered further in this chapter. In the following sections the method of Southern blotting, which reveals restriction site variation in DNA, and the technique of DNA sequencing will be emphasized.

Southern Blotting and Inferences from Restriction Site Variation

One widely used method for examining molecular sequence variation is the Southern blot. In this method, DNA samples from many individuals are each digested with restriction enzymes. The resulting fragments are separated by size using electrophoresis, and are then transferred ("blotted") to a nylon or nitrocellulose filter. Next the filters are mixed with radioactive probe DNA under conditions in which the probe hybridizes with homologous restriction fragments on the filter; migration distances, and thus the size of the fragments, are determined by autoradiography. In restriction site analysis using Southern blots, variation among alleles is detected as fragment size differences due to either (1) presence or absence of restriction sites or (2) insertion or deletion variation. These possibilities can be distinguished by digesting the DNA with more than one restriction enzyme, followed by analysis of fragment sizes. The sizes of insertions and deletions can be estimated by the change in the migration distance of the bands in the gel. For fragments smaller than about 20,000 base pairs (20,000 bp, or 20 kb) the migration distance is roughly proportional to the logarithm of the length. A mathematical function that fits the relationship between DNA fragment size and electrophoretic migration even better is

$$(m_i - m_0)(L_i - L_0) = c \qquad (7.1)$$

where L_i is the length of fragment i, m_i is the migration distance for fragment i, and c, m_0 and L_0 are constants. In practice, a standard series of known size standards is run with the samples in the electrophoresis gel. Least-squares (regression) methods are used to fit Equation 7.1 to the size standards, allowing estimation of unknown fragment sizes from their migration distance in the gel (Figure 1).

Although the data from Southern blot analysis allow one to determine which individuals contain the same alleles with respect to the particular sites examined, it requires some mathematics to go from those data to estimates of heterozygosity of nucleotide sites and levels of polymorphism. The necessary manipulations are as follows.

First, consider the situation where the location of restriction sites (restriction map) of the gene is unknown, but data are available on the numbers

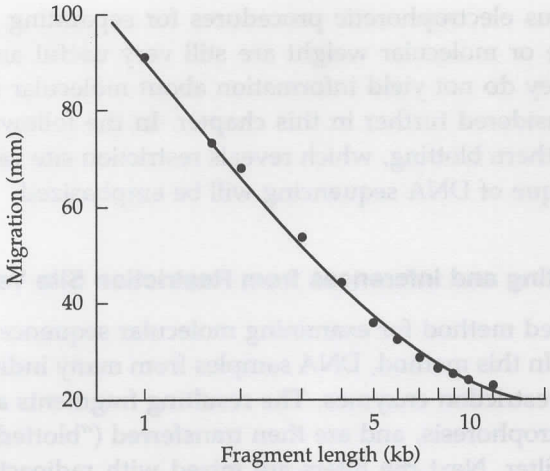


Figure 1. The relationship between fragment length and migration distance on an agarose gel. Fragment sizes can be estimated by running standards having known fragment lengths in another lane of the same gel. For fragments that are longer than about 20,000 bp, other methods (such as field inversion electrophoresis) give better separation and size estimates.

and sizes of restriction fragments from several individuals or several species. The methods of Upholt (1977), Nei and Li (1979), and Engels (1981) all deal with this situation, and we now consider Engels' method in some detail. The objective is to obtain estimates of the level of nucleotide polymorphism (the fraction of nucleotide sites that vary among individuals in a population) and heterozygosity (the chance that a random position in the sequences of a random individual will differ in the two homologs). Consider the segment of DNA as having m restriction sites, with the number of sequences in the sample that are cut at each site being $c_1, c_2, \ldots c_m$. In a sample of size n, these values are constrained within $0 \leq c_i \leq n$. Let c represent the total number of cuts at all restriction sites. If k of the sites are polymorphic (that is, in some of the sampled individuals the site is cut, but in others it is not, implying that any of the j positions in the restriction site is polymorphic), then k/m represents the fraction of sites that are polymorphic. This ratio is not a useful measure of polymorphism, however, because it is strongly biased in that the observation of a polymorphic restriction site is conditional on having observed a cut at the site at least once in the sample. However, if we assume that the frequency of a recognition sequence depends only on the nucleotide frequencies, and that the probability of a monomorphic site does not depend on the nucleotide sequence, then the expected number of monomorphic sites is

$$\mathrm{E}(m - k) = (L - j + 1)\ \Pr(\text{monomorphic site})\ \Pr(\text{recognition sequence}) \quad (7.2)$$

where L is the length of the segment analyzed, j is the length of the recognition sequence, and Pr(recognition sequence) is the probability of obtaining a recognition sequence by chance in a sequence of length L with the observed nucleotide distribution. Although one could derive Pr(recognition sequence) after making assumptions about the distribution of nucleotides (complete independence would lead to a multinomial distribution), it is not necessary for estimating the fraction P of nucleotide sites that are polymorphic. The probability of a given site being monomorphic can be estimated as the expected number of monomorphic sites per sequence $\mathrm{E}(m - k)$, times the number of sequences in the sample (n), divided by the expected total number of cuts $\mathrm{E}(c)$:

$$\Pr(\text{monomorphic site}) = n\ \mathrm{E}(m - k)/\mathrm{E}(c) \quad (7.3)$$

To estimate this quantity, the expected values can be replaced by the observed values, so that $\langle \Pr(\text{monomorphic site}) \rangle = n(m - k)/c$. The probability of a polymorphic site is 1 − Pr(monomorphic site), or $[c - n(m - k)]/c$. Assuming that no more than one of the nucleotides in a restriction site is polymorphic in the sample (Ewens et al. 1981), then

$$P = \Pr(\text{polymorphic restriction site})/j \quad (7.4)$$

In the case of a four-cutter restriction enzyme (that is, one with a four-nucleotide restriction site), in a population in which 20% of the restriction sites are polymorphic, the individual nucleotides will be polymorphic an average of 20/4 = 5% of the time. Combining Equation 7.4 with an estimate of the probability of a polymorphic site derived from Equation 7.2, we obtain

$$\langle P \rangle = \frac{c - n(m - k)}{jc} \quad (7.5)$$

as an estimate of the fraction of nucleotide sites that are polymorphic.

To estimate the heterozygosity of nucleotide sites within a population from data on restriction site variation, we assume that

$$H = \Pr(\text{heterozygous site})/j \quad (7.6)$$

just as in Equation 7.4. Using statistical methods beyond the scope of this book, Engels (1981) derives

$$\langle H \rangle = \frac{nc - \Sigma c_i^2}{jc(n - 1)} \quad (7.7)$$

Equation 7.7 provides a means of estimating the heterozygosity per nucleotide, by Southern blot analysis of restriction site variation, of individuals from a natural population.

Nei and Li (1979) have considered the problem of estimating the degree of evolutionary divergence of DNA sequences as ascertained by restriction site variation. When F is the fraction of DNA restriction fragments that are shared by two species (or populations) and M is the probability that a restriction site has remained unchanged, they show that

$$F \approx M^4/(3 - 2M) \tag{7.8}$$

F can be estimated from a population sample using

$$\langle F \rangle = \frac{2n_{XY}}{(n_X + n_Y)} \tag{7.9}$$

where n_X and n_Y are the numbers of DNA fragments in the digests from species X and Y, and n_{XY} is the number of fragments shared by the two species (assuming no insertion/deletion variation). To estimate M, Equation 7.8 can be rearranged to yield

$$\langle M \rangle = [\langle F \rangle (3 - 2M)]^{1/4} \tag{7.10}$$

which can be solved by iteration, where successive values on the left side are calculated using the previous value substituted on the right side until both sides agree. The first trial value of M can be taken as $\langle F \rangle^{1/4}$ (Nei 1987).

As will be shown later in this chapter, the probability that a restriction site remains unchanged in time t is $M = e^{-r\lambda t}$, where r is the number of nucleotides per restriction site, and λ is the rate of nucleotide substitution. Furthermore, the expected number d of nucleotide substitutions per site is $d = 2\lambda t$. Therefore, the estimate of $\langle M \rangle$ allows us to get an estimate of d, which is a measure of evolutionary divergence, as follows:

$$\langle d \rangle = -(2/r)\ln(\langle M \rangle) \tag{7.11}$$

As an example, suppose a pair of species is compared using restriction fragment length variation detected by Southern blot analysis with five six-cutter enzymes, and that $n_X = 30$, $n_Y = 20$ and $n_{XY} = 15$. By applying Equation 7.9, we get $\langle F \rangle = (2)(15)/(30 + 20) = 0.6$, and the trial value $M = F^{1/4}$ leads, after 20 rounds of iteration of Equation 7.10, to $M = 0.915$. Hence, since $r = 6$ in this case, the estimate of $\langle d \rangle$ is 0.030. This is one method for estimating the degree of divergence of DNA based only on DNA fragment lengths, and Kaplan's (1983) computer simulations showed it to be quite reliable as long as $d < 0.05$. For more divergent sequences, this method underestimates the degree of divergence.

More reliable estimates of the proportion of polymorphic sites and the number of nucleotide substitutions per site can be obtained when the restriction map is known, allowing inference of which sites are monomorphic and which vary. Means of estimating population parameters with this kind of data have been explored by Kaplan and Langley (1979), Kaplan and Risko (1981), Gotoh et al. (1979), and Nei and Tajima (1983). The latter used a maximum likelihood estimation method that also allowed estimation of sampling variances of the statistics. The basic idea of the method is to express the likelihood of obtaining the observed data given n_X, n_Y, and n_{XY} (defined following Equation 7.4), and m_T, the total number of nucleotides examined. The likelihood is a function of a, the probability that a randomly chosen set of r contiguous sites is a restriction site; and S, the probability that a pair of sequences X and Y share a given recognition site. This likelihood is maximized at a value that yields the maximum likelihood estimators for a and S, which are

$$\langle a \rangle = \frac{n_X + n_Y - n_{XY}}{m_T(2 - S)} \tag{7.12}$$

$$\langle S \rangle = \frac{2n_{XY}}{n_X + n_Y}$$

With these estimates in hand, the proportion of polymorphic nucleotide sites and the number of substitutions per site are estimated as

$$\langle P \rangle = 1 - \langle S \rangle^{(1/r)}$$
$$\langle d \rangle = [-\ln(\langle S \rangle)]/r \tag{7.13}$$

The best way to appreciate the advantage of knowing the restriction map information is to examine some examples where variation among individuals within a species could be located in the restriction map. In the case of *Drosophila*, such studies have revealed not only site differences, but also a surprising level of length variation (Box A).

Restriction Map Variation in the *Adh* Gene of *Drosophila melanogaster*

A

Using DNA extracted from 48 lines of *Drosophila melanogaster* captured in four regions of the United States, Aquadro et al. (1986) performed a detailed study of restriction map variation of a 13-kb region of the second chromosome containing the *Adh* gene. The DNA samples were digested with the six-cutters *Bam*HI, *Bgl*II, *Ban*II, *Eco*RI, *Hind*III, *Hha*I, *Sal*I, and *Xho*I, and fragments were separated by agarose gel electrophoresis,

blotted onto nitrocellulose, and visualized by probing with the cloned *Drosophila Adh* gene. The data are summarized in Chapter 1 (Figure 11), where monomorphic and polymorphic restriction sites are shown, as well as a diverse set of insertion and deletion variants. The variation reveals that $\langle P \rangle = 0.0270$, implying that about 1 in 37 nucleotides is polymorphic; and $\langle H \rangle = 0.0064$, implying that about 1 in 156 nucleotides differ between any two randomly chosen chromosomes.

The variation in frequency of insertion and deletion variation was so great as to require reevaluation of the relative magnitudes of point mutations and mutational events in changing restriction fragment lengths. For example, 80% of the lines had one or more fragments differing in length from the consensus restriction map. Identification of the sequences inserted was made by hybridization using a series of previously cloned *Drosophila* transposable elements (described later in the chapter). The elements *copia*, B*104*, *2161*, and *101*F were all found to have inserted into the *Adh* region in one or more of the lines. There were also a number of unique length variants, ranging in size from 21 to 200 bp.

Among the 48 lines examined, there were 29 distinguishable types of chromosomes (haplotypes). The haplotype diversity, or virtual heterozygosity (defined as $1 - \Sigma p_i^2$ where p_i is the frequency of the ith haplotype), shows an interesting pattern when separated into sequences that correspond to the *fast* and the *slow* allozyme forms. The *slow* had a haplotype diversity of restriction sites of 0.819, and the *fast* alleles had a much lower restriction site diversity of 0.419. Indeed, 75% of the *fast* alleles had the same restriction site haplotype, which means that there was extensive linkage disequilibrium among the sites. The decreased diversity of the *fast* alleles, along with the genealogy of the alleles, suggests that the *fast* allele is younger (more recently diverged) than the *slow* allele. Interestingly, the *fast* allele has a greater diversity for transposable element insertion (0.507) compared with the *slow* allele (0.339).

The *Adh* gene of *Drosophila melanogaster* has received a great deal of attention, and we will return to analysis of sequence variation among alleles. Kreitman and Aguade (1986a,b) used a refinement of the Southern blot to greatly increase its resolution. Using four-cutter enzymes that cut the DNA in many places, many small fragments are produced that differ only slightly in size. These are separated on a long, thin DNA sequencing gel, then blotted onto a nylon filter and probed as before. With this technique it is possible to separate DNA fragments that differ in size by a single nucleotide (Figure 2). The high-resolution analyses generally confirmed the size variation summarized in Box A, and revealed that populations as far apart as Raleigh, North Carolina and Putah Creek, California, had similar haplotypes.

Restriction site variation in the region of the *Adh* gene has also been examined in *Drosophila pseudoobscura* (Schaeffer et al. 1987). In contrast to the strong linkage disequilibrium, high level of insertion/deletion variation, and relatively low level of nucleotide polymorphism found in *D. melanogaster*, *D. pseudoobscura* exhibited a higher level of polymorphism ($\langle P \rangle = 0.066$) and

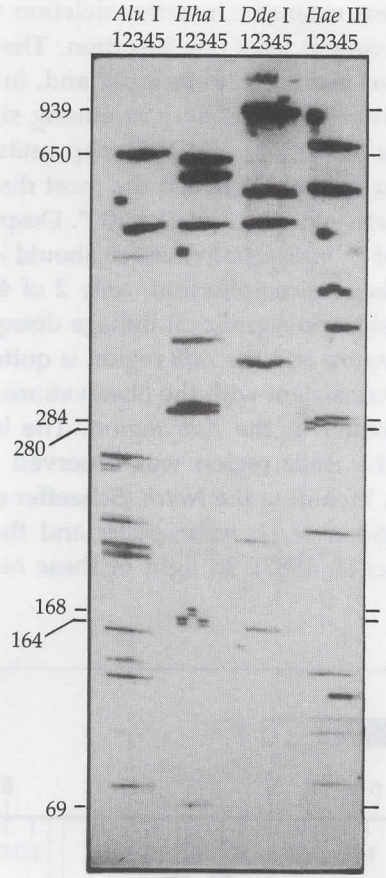


Figure 2. Autoradiograph of a Southern blot of five *Drosophila melanogaster* DNA samples digested with four restriction enzymes and probed with a 2.7-kb homologous probe encompassing the *Adh* locus. Large polyacrylamide sequencing gels have sufficient resolution to distinguish DNA fragments that differ in length by a single nucleotide. The range of fragment sizes in bp is shown at the left. A 1-bp insertion/deletion difference between samples 2 and 3 can be seen in the *Hae*III (283–284 bp) fragments. (From Kreitman and Aguadé 1986a.)

heterozygosity ($\langle H \rangle = 0.021$), and only 2 of 20 lines showed insertion/deletion differences. Finally, there was high haplotype diversity and low linkage disequilibrium in *D. pseudoobscura*, suggesting older polymorphisms or lack of population bottlenecks, or both. The interspecific comparison is taken to the nucleotide sequence level in Box D.

Restriction map variation has also been extensively examined among 38 lines of *D. melanogaster* in the region of the *white* gene (Langley and Aquadro

1987). Here again there was extensive insertion/deletion variation (Figure 3), with 37% of the lines having at least one insertion. The level of nucleotide diversity in this region was estimated to be 0.012 and, in contrast to the *Adh* region, there was little linkage disequilibrium among sites. The rate of recombination in *Drosophila* is roughly 2×10^{-5} map units per 1 kb of DNA, so the effective recombination rate between the most distant sites examined in the *white* region was between 10^{-5} and 2×10^{-4}. Despite the tight linkage found in the *Adh* region of *D. melanogaster*, which should increase the chances of finding significant linkage disequilibrium, only 2 of 49 possible pairwise tests in the *white* region showed significant linkage disequilibrium. The contrast between the *white* region and the *Adh* region is quite remarkable in this regard. One explanation consistent with the observations is that the intensity of natural selection is greater in the *Adh* region. The low level of linkage disequilibrium seen in the *white* region was observed in other studies of restriction map variation, including the *Notch* (Schaeffer et al. 1988) and *Amy* (Langley et al. 1988) regions of *D. melanogaster* and the *Adh* region of *D. pseudoobscura* (Schaeffer et al. 1987). In light of these other observations, it

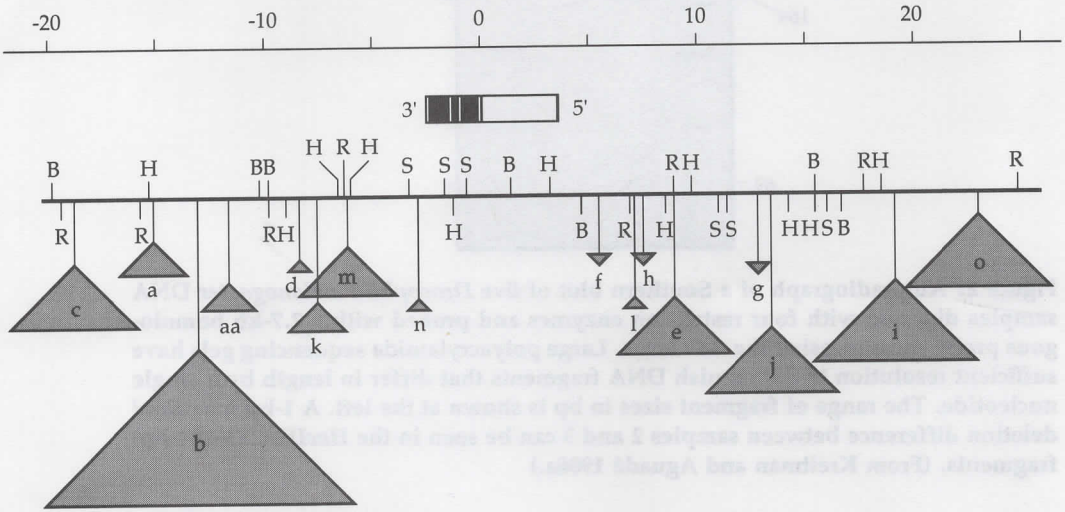


Figure 3. Restriction map variability in the *white*-locus region of *Drosophila melanogaster*. The numbered line at the top presents the map coordinates (in kilobase pairs), and the rectangle represents the location of the *white* gene. Monomorphic restriction sites are marked above the heavy horizontal line, and polymorphic restriction sites, insertions, and deletions are marked below the line (B = *Bam*HI; H = *Hind*III; R = *Eco*RI; and S = *Sal*I). The strain *Oregon-R* has the most common haplotype, and it serves as the standard in the figure. Sizes of the large insertions are to scale in the figure. The smaller insertions and deletions vary in size and are indicated by lower case letters. (From Langley and Aquadro 1987.)

appears that the *Adh* region in *D. melanogaster* is exceptional and that it has an unusual evolutionary history.

The elucidation of restriction site variation is particularly useful in humans, especially when the gene being probed is one that may be responsible for genetic disorders. The utility of a population genetic approach to this problem was first demonstrated by Kan and Dozy (1978), who found strong linkage disequilibrium between the $Hb\beta^S$ allele and a particular restriction fragment. Tight linkage between RFLP markers and genes responsible for genetic disorders is of great value in prenatal risk estimation. The globin region is rich in lesions that lead to a variety of hemoglobinopathies, and these simple molecular techniques have directly demonstrated the cause of many of them. There are, for example, over 30 different mutations that lead to β-thalassemia, involving transcription, RNA processing, and translation (Orkin and Kazazian 1984).

DNA Sequencing

The ultimate means of identifying a gene and resolving allelic differences is to determine the nucleotide sequences directly. Great strides have been made in simplifying this laborious procedure, and the impact of this technology on evolutionary genetics is just beginning. Molecular biologists often recognize remarkable evolutionary relationships between genes simply because of their DNA sequence similarity. For example, flies containing the mutation *technical knockout* (*tko*) in *Drosophila melanogaster* remain motionless for periods of up to a minute when the culture vial is tapped, whereas wildtype flies get irritated and buzz around. After cloning and sequencing a region of the genome known to contain the *tko* gene, Royden et al. (1987) observed a remarkable sequence similarity to ribosomal protein S12 in *Escherichia coli* and to ribosomal proteins in the chloroplast of *Euglena gracilis*. The significance of the similarity has not been elucidated yet, but the sequence match alone provides a clear direction for the molecular biologists to pursue. As the molecular sequence databases grow and the computer software for performing massive searches gains sophistication, it is certain that this approach will become easier, more routine, and even more powerful.

DNA sequence data are qualitatively different from protein electrophoretic data in that the problem of undetected variants does not arise. Moreover, the relative levels of sequence variation in different parts of genes and their flanking regions suggest the modes of evolutionary change that shaped the genes. The power of DNA sequence data is shown by an example (Figure 4) in which 11 alleles were sequenced in a 2659-bp region of the *Drosophila melanogaster* genome including the *Adh* locus (Kreitman 1983). Kreitman used

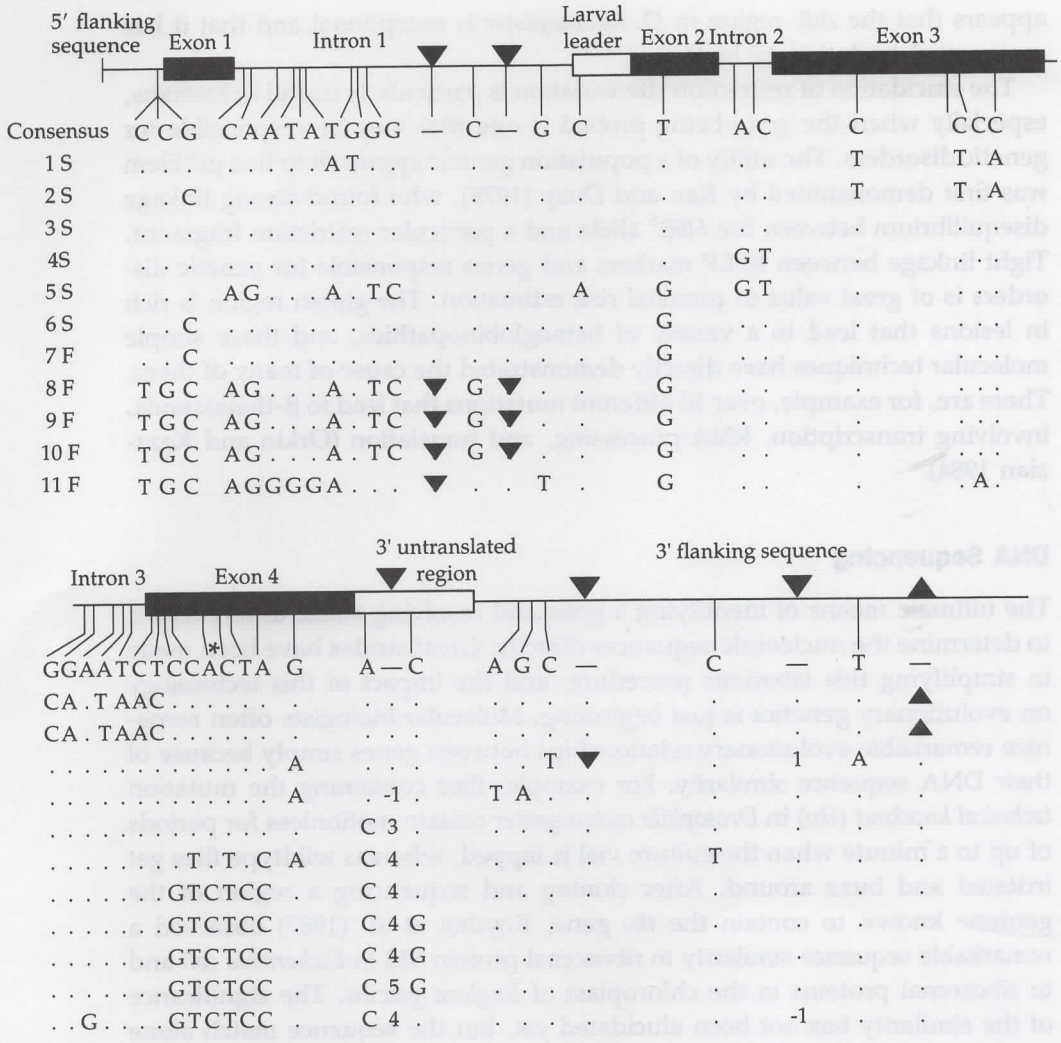


Figure 4. Polymorphic nucleotide sites among 11 alleles of the alcohol dehydrogenase gene of *Drosophila melanogaster*. The first line gives a consensus sequence for *Adh* at sites that vary; subsequent lines give the nucleotides from each copy for the polymorphic sites. A dot indicates that the site is identical to the consensus sequence. The triangles indicate sites of insertion or deletion relative to the consensus sequence. The star in exon 4 indicates the site of the amino acid replacement (lysine for threonine) responsible for the *fast-slow* mobility difference in the *Adh* protein. (After Kreitman 1983.)

the sequencing method of Maxam and Gilbert (1977) outlined in Box B of Chapter 1. Another commonly used method of DNA sequencing is presented in Box B of this chapter.

Sanger Dideoxy Chain-Termination Method for Sequencing DNA

B

Although the Maxam and Gilbert (1977) method is still used in many laboratories (see Box B in Chapter 1), the dideoxy chain-termination method of Sanger et al. (1977) is, for most purposes, simpler, faster, and more accurate. The dideoxy method requires single-stranded DNA, which can be obtained by using a single-stranded cloning vector, such as the bacteriophage M13. The principle behind the method is that the large fragment of DNA polymerase (the Klenow fragment) is able to synthesize a radioactively labeled strand complementary to the DNA template to be sequenced. In order for the

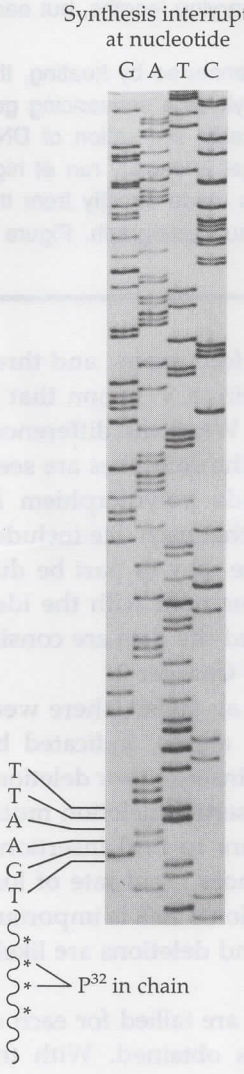


Figure 5. In the Sanger sequencing method, four aliquots containing a single-stranded DNA template are replicated in the presence of all four normal nucleotides plus a small amount of one of the nucleotides in the dideoxy form. Dideoxynucleotides lack an oxygen atom required for attachment of the next nucleotide in line, so synthesis stops whenever the replicating chain incorporates a dideoxy nucleotide. When the resulting fragments are separated by length in a gel, the nucleotide present at a particular position in the replicating chain is indicated by which of the four dideoxynucleotides caused termination of synthesis at that position. The diagram at left shows how the sequence TGAATT is read from the gel. The asterisks indicate radioactivity present in the DNA chain from using at least one of the four normal nucleotides in radioactive form. The first 50 nucleotides in this particular sequence are CTCATTGAATTAAATCTCTTGTGT GTAAATTCTACTCATTGCAGGTGAAA. (Courtesy of R. F. DuBose.)

synthesis to begin, the polymerase must also be provided with a short, single-stranded primer. The M13 system offers the additional advantage of having commercially available universal primers that initiate the reaction within M13 DNA. Four separate reactions are run, one for each nucleotide. In each reaction mixture there is the template, the primer, the polymerase, and the four deoxynucleotides (ATP, GTP, CTP, and TTP), at least one of which is radioactive. In addition, each reaction mixture contains the dideoxy form of one of the four nucleotides. In each reaction, when the polymerase incorporates a dideoxynucleotide into the growing strand, the synthesis terminates because the dideoxynucleotide is missing the 3′ hydroxyl group required for formation of the next phosphodiester bond in the DNA backbone. Since each reaction mixture contains both deoxy and dideoxynucleotides, the incorporation of the dideoxy form occurs at random, which results in random termination of the chains. For example, the dideoxy-ATP reaction mixture will generate a series of DNA fragments of varying lengths, but each fragment terminates with an A.

After the reactions are stopped and the DNA has been denatured by heating, the contents of each reaction mixture are loaded onto a polyacrylamide sequencing gel. Sequencing gels are generally quite long (40–80 cm) to enable separation of DNA fragments that differ in length by only one nucleotide. The gel is usually run at high voltage for about six hours, then dried; an autoradiograph is made directly from the dried gel. Each DNA fragment in the gel shows up in the autoradiograph. Figure 5 shows the method and an example.

Kreitman's (1983) data include a 5′ flanking region, four exons, and three introns of the larval and adult forms of the *Adh* protein, a 3′ region that is transcribed but untranslated, and a 3′ flanking region. When the differences among the sequences are aligned as in Figure 4, 43 of the 2659 sites are seen to differ, so 43/2659 (0.0162) is the level of nucleotide polymorphism P. Another important feature of the polymorphic sites is that each site includes only two nucleotides, never three or four. This feature may in part be due to the small number of alleles sequenced, but it is consistent with the idea that mutations at a particular site are rare. In this regard the data are consistent with the formulation of the infinite sites model in Chapter 2.

As in the restriction fragment study by Aquadro et al. (1986), there were a number of insertions and deletions among the 11 alleles, indicated by triangles in Figure 4. As might be expected, none of the insertions or deletions are found within the coding region. Assuming that insertion/deletion mutations occur randomly across the gene, then the failure to find insertions/deletions in the coding regions may be due to their more rapid rate of loss from the population. This would be expected if a functional *Adh* is important for the viability and fertility of a fly, since insertions and deletions are likely to disrupt function.

When the proportions of sites that are polymorphic are tallied for each of the parts of the *Adh* region, the plot in Figure 6 is obtained. With the

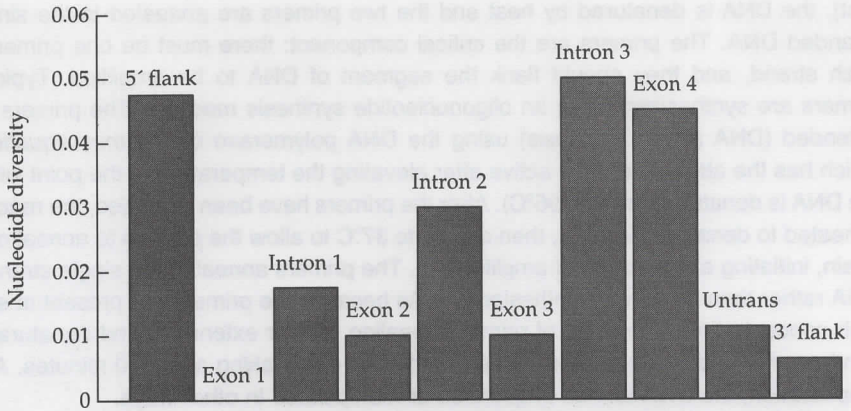


Figure 6. Nucleotide diversity in the *Adh* gene of *Drosophila melanogaster*.

exception of the highly polymorphic fourth exon, there is a clear tendency for the exons to be less variable than the introns and 5′ flanking region. The unexpected observation of a conserved 3′ flanking region remains unexplained. Of the 14 polymorphic sites in the coding region, 13 result in no change in the amino acid sequence of the gene product. If mutations were random, nearly twice as many amino acid replacements as silent substitutions would occur. In fact, silent substitutions outnumber amino acid replacements, which suggests that selection eliminates most mutations that result in amino acid replacements. Sequences of *Adh* in *Drosophila pseudoobscura*, *D. mauritiana*, *D. simulans*, and *D. orena* have also been determined (Schaeffer and Aquadro 1987; Cohn et al. 1984; Bodmer and Ashburner 1984). When compared with the *Adh* of *D. melanogaster*, some peculiar patterns emerge. For example, the silent sites are nonrandomly distributed across exons, suggesting that there may be selective constraints acting even on the silent substitutions. The problem of inferring the action of natural selection and selective constraints from sequence data will be considered in more detail later in this chapter (page 376).

The Polymerase Chain-Reaction Method of DNA Amplification

C

Since DNA polymerase makes copies of single-stranded DNA when provided with a primer, in principle it should be able to make copies of the copies if the DNA is repeatedly denatured and short primers are available for both strands. This is the principle behind the polymerase chain reaction (Saiki et al. 1985; Scharf et al. 1986; Engelke et al. 1988) and, as the name implies, it can generate a virtually unlimited number of copies of a particular DNA sequence. Starting with unamplified DNA (for example, DNA from a hair

root), the DNA is denatured by heat and the two primers are annealed to the single-stranded DNA. The primers are the critical component: there must be one primer for each strand, and they should flank the segment of DNA to be amplified. Typically primers are synthesized using an oligonucleotide synthesis machine. The primers are extended (DNA strand synthesis) using the DNA polymerase of *Thermus aquaticus*, which has the ability to remain active after elevating the temperature to the point where the DNA is denatured (around 95°C). After the primers have been extended, the reaction is heated to denature the DNA, then cooled to 37°C to allow the primers to anneal once again, initiating a new round of amplification. The primers anneal to the single-stranded DNA rather than the newly synthesized DNAs because the primers are present at such high concentration. The cycle of primer annealing, primer extension, and denaturation can be continued for 30 or more cycles, with each cycle taking about 10 minutes. After amplification, the DNA can be sequenced or manipulated in other ways.

The polymerase chain reaction (PCR) amplifies segments of DNA to very high copy numbers — $2^{30} \approx 10^9$ in the case of 30 cycles. The method is becoming very popular in molecular population genetics because it eliminates the need to separately clone each allele to be sequenced. Once any sequence has been determined, oligonucleotide primers can be synthesized to examine sequence variation at the population level by amplifying the DNA from a series of individuals. Without PCR it would be necessary to create and isolate DNA clones from each sample, greatly increasing the time and effort. One problem is that the amplified DNA from a heterozygote that contains two different sequences for the two alleles generates ambiguity in the sequencing gels. For example, if a particular site had an A in one homologous chromosome and a T in the other, this would appear as both an A and a T on the sequencing gel. Insertions and deletions cause even more ambiguity. One way to avoid this problem is to use genetic methods to get homozygous individuals — for example, by using balancer chromosomes in *Drosophila*, or one can use haploid tissue such as the gametophytes of plants.

DNA–DNA Hybridization

In DNA hybridization, DNA molecules from two organisms are separated into individual strands at high temperature and then mixed to promote reannealing of the strands to form heteroduplex molecules that contain one strand from each organism. The heteroduplex molecules normally contain nucleotide mismatches, and the degree of mismatch is measured by the difference in melting temperature (ΔT_m), defined as the difference between the temperature at which half of the heteroduplex molecules denature into their individual strands and the temperature at which half of the homoduplex molecules denature. In practice, the DNA from one organism is rendered radioactive, and the melting temperature is determined from the temperature at which radioactive strands are released from the heteroduplex molecules. As a rule of thumb, a change in ΔT_m of 1°C results from 1% mismatches in

the heteroduplex molecules. The method of DNA–DNA hybridization has been used for inference of phylogenetic relations among many taxa, including birds (Sibley and Ahlquist 1981) and primates (Sibley and Ahlquist 1984, 1987). Applications of the method will be discussed later in this chapter in relation to the construction of phylogenetic trees. An important assumption made in constructing phylogenies from DNA–DNA hybridization is that the value of ΔT_m is linearly related to the proportion of nucleotide difference between the two DNA sequences. Empirical data support this approximation.

DNA Sequence Alignment Methods

D

Once a DNA sequence has been determined, evolutionary inferences generally require some sort of alignment among sequences, followed by a test of similarity to the other sequences. The other sequences might be alternative alleles, as in Figure 4, or sequences from other species. Alignment with sequences from other species is central to the inference of phylogeny, or ancestral history, from molecular data.

There are two reasons to obtain sequence alignments. The first is to determine whether there is significant similarity among a set of sequences, so that inferences about whether they are homologous (share a common ancestor) can be made. The other reason is to enable analysis such as that in Figure 4, which requires a base-by-base comparison of homologous sequences to determine the degree and structure of differentiation.

Many commonly used methods for aligning sequences are based on the Needleman-Wunsch algorithm (Needleman and Wunsch 1970). The basic idea is to imagine the two sequences written out as column and row headings in a matrix. Any alignment of the sequences can be depicted as a path through the matrix: a straight diagonal line represents no insertion/deletion variation, while zig-zags represent insertions and deletions in the alignment. As any alignment is traced through the cells of the matrix, every cell in the matrix (i, j), is assigned a score according to the three cells that could precede it: $(i - 1, j - 1)$ if along the diagonal, $(i, j - 1)$ corresponding to a vertical move and a lack of match in the first sequence, and $(i - 1, j)$ corresponding to a horizontal move and a lack of match in the second sequence. For any path through the matrix, the score of the path is obtained by summing the matches (diagonal steps) and subtracting the number of steps that represent insertion/deletion events, each multiplied by a "gap penalty." The highest such score represents the "best" alignment. However, in practice there are several ways to define a best alignment, including a weighting for the best alignment in some defined region, or for the best overall sequence alignment. Efficiency is improved in computer algorithms that do alignments by ignoring some elements of the score matrix. For example, elements far from the diagonal are often not useful. Multiple sequences can be aligned by viewing them as axes in a multidimensional lattice, and the best sequence is a path through this lattice. There is no single "best" way to define an optimal alignment, since it depends on the gap penalties, and the arbitrariness of the definition should be kept in mind when examining sequence alignments.

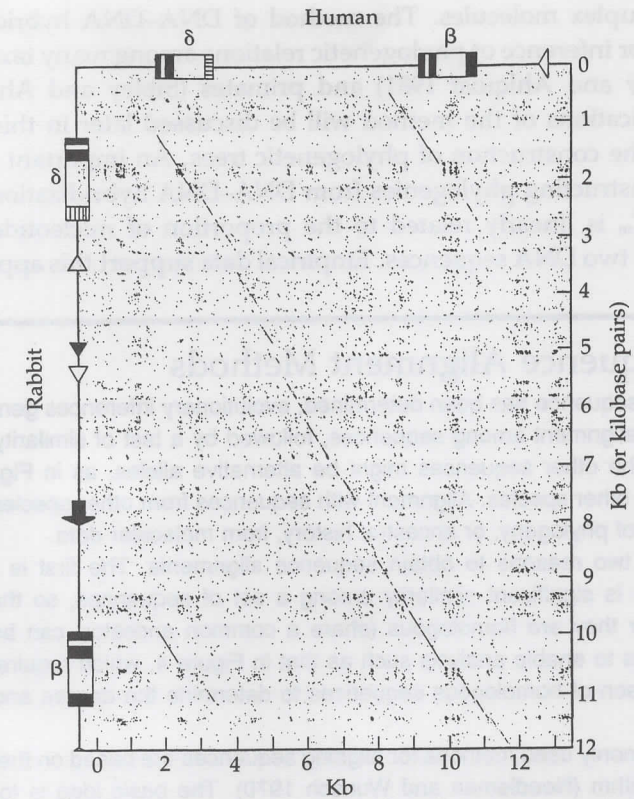


Figure 7. A dot plot comparison of the human and rabbit sequences spanning δ- and β-globins. The scales are in kilobases and the rectangles indicate the location and organization of the globin genes. The solid arrows show the location of a rabbit L1 repeat, and open triangles indicate human *Alu* sequences and rabbit *OcC* repeats (a rabbit *SINE*). The major diagonal line indicates that there is noticeable homology retained through the δ-β intergenic region, and the sequence similarity of human β-globin with rabbit δ-globin (and vice versa) is evident. (From Margot et al. 1988.)

Another approach to alignment is to search for sequences or parts of sequences that show similarity. A graphical way to do this is with a **dot plot**. The two sequences are written out along the x and y axes, and a "window" of size w and "minimal match criterion" c are selected. For every point (i,j) in the plane, the bases i to $i + w - 1$ in sequence 1 and bases j through $j + w - 1$ in sequence 2 are compared. The number of bases in the sequences that match are counted, and if the matches exceed c, then a diagonal line from (i,j) to $(i + w, j + w)$ is plotted. When all entries (i, j) are considered in turn, something like Figure 7 is obtained. Sequence similarity is judged by the occurrence of diagonal lines and by the stringency of the match (determined by c/w). Dot plot construction is quite computation-intensive: when the sequence lengths increase by a factor of x, the computer time needed to construct the plot grows as x^2.

A similar but quicker method of alignment only considers exact matches of length w

or longer (i.e., $c = w$). The probability distribution for finding the longest such match (called an **identity block**) has been determined by Karlin et al. (1988) in a paper that also describes many other statistical properties of identity blocks. Identity blocks can be found very quickly using a computing technique known as **hashing**. If the window size is 4 (typically it is 10 or more), there are $4^4 = 16$ possible sequences of length 4, which are called hash bins. The computer program sweeps through each sequence and records the start position of each block that matches a hash bin. Once this is done for the two sequences, a "block" plot can be drawn, and counts of the number of matches are easily tallied. An example of a block plot appears later in the chapter (Figure 28), and is discussed further in the section on multigene families and ribosomal RNA. Note that the time required for hashing sequences is proportional to length, not the square of the length. This allows block identity algorithms to quickly find, for example, that the longest exact match in the human, mouse, bovine, and *Xenopus* mtDNA sequences occurs in the 16S rRNA gene and is 30 bp long (Karlin 1986). Because of this tremendous saving in computer time, hashing is often used for matching a given sequence with those in a large database.

Microcomplement Fixation

One method that has proven useful for obtaining estimates of evolutionary divergence because of its relative speed and simplicity is microcomplement fixation, first introduced by Sarich and Wilson (1966). The method is of historical importance because the results directly opposed the traditional view of paleontologists that humans belong in a separate family and that the divergence from chimpanzees and gorillas occurred 20 million years ago. Rather, Sarich and Wilson (1966, 1967) found that chimpanzees and gorillas are much more closely related to humans, and the split occurred about 5 million years ago. Paleontological data collected since that time have largely corroborated Sarich and Wilson's findings.

Microcomplement fixation has been widely applied in the analysis of anuran (toads and frogs) systematics, and has revealed extensive molecular divergence with relatively little morphological change. Particularly noteworthy is the poor correspondence between the magnitude of morphological change and of molecular divergence in this group (Maxson 1984).

Microcomplement fixation is used to determine the degree to which the antigenic specificity of proteins has diverged, and is based on the thesis that the degree of antibody binding decreases as proteins become more divergent (Champion et al. 1974). It has proven useful because, for small degrees of divergence, microcomplement fixation shows a high correlation with amino acid sequence variation (Benjamin et al. 1984).

The test uses antisera containing antibodies against a specific protein, such

as albumin. The test itself is carried out by mixing the antiserum with each antigen at a series of concentrations in the presence of serum proteins known as "complement." Among a set of antigens, the more divergent antigens elicit a weaker antibody–antigen reaction. The strength of the reaction is measured according to the percentage of the complement protein that is bound to the complexed antibody (fixed). More complement fixation means a closer antigenic match. Figure 8 illustrates how the immunological distance (ID) between antigens is calculated from the test. The ID statistic has been shown to be approximately linearly related to the time of divergence, provided that the divergence times are not too large (Sarich and Wilson 1967).

After all pairwise tests have been conducted on a set of taxa, the resulting matrix of IDs can be used to construct a phylogeny. The limitations of such phylogenies are very similar to those constructed using DNA–DNA hybridization because the data are single summary values. Use of microcomplement fixation has been considerably strengthened by the observation that ID is

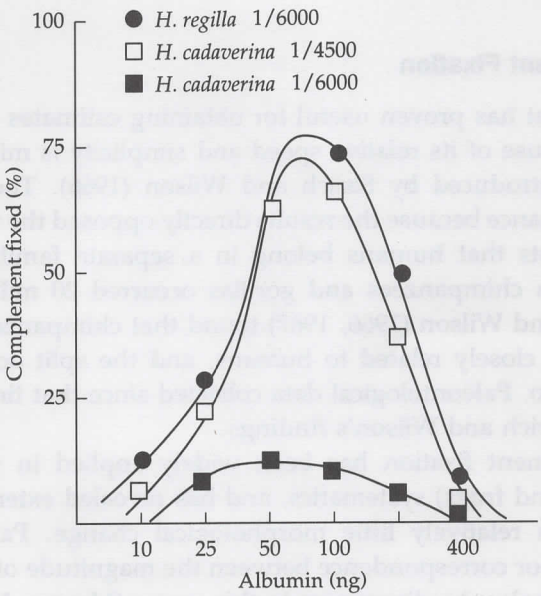


Figure 8. The method of microcomplement fixation is based on the strength of immunological crossreactivity of proteins as assessed by these microcomplement fixation (MCF) curves. In this example the antiserum is to *Hyla regilla* albumin at a relative concentration of 1/6000 (solid symbols) and 1/4500 (open symbols). The albumin is either that of *H. regilla* (solid dots) or of *H. cadaverina* (open and solid squares). Since the heterologous curve (open squares) at a 1/4500 dilution fixes complement equivalent to the homologous curve at 1/6000 (solid dot), the immunological distance between these two species is defined as 100log(6000/4500) = 12.5. (From Maxson and Maxson, 1986.)

related to the number of amino acid replacements. According to a molecular model of Maxson and Maxson (1986), each antigen has a large number of antigenic sites, only a small minority of which bind antibody. In a divergent antigen, the same degree of complement fixation can be achieved only if additional antiserum is added. The reason is that, while the divergent antigen is lacking some of the highly specific sites, the remaining sites can still bind antibodies. The number of highly specific sites that are missing is thus linearly related to the amount of additional antiserum necessary to achieve the same complement fixation. This model, called the assortment–exclusion hypothesis, is rather well corroborated by direct experimental test (Maxson and Maxson 1986).

THE NEUTRAL THEORY AND MOLECULAR EVOLUTION

The first systematic application of molecular methods to population genetics in the 1960s, initially using the electrophoretic separation of protein molecules, revealed the existence of extensive genetic variation within most natural populations. In most species 15–50% of the genes coding for enzymes were observed to include two or more widespread polymorphic alleles. The polymorphic alleles occurred with frequencies considered to be too high to result from equilibrium between adverse selection and mutation.

In 1968, Motoo Kimura suggested that most polymorphisms observed at the molecular level are selectively neutral (Kimura 1968). Mutations that are selectively neutral produce such small effects on the ability of their carriers to survive and reproduce that they are completely equivalent in terms of natural selection. The ultimate fate of neutral alleles in a population is determined largely by the process of random genetic drift. By extension, the hypothesis of selective neutrality would also apply to most nucleotide or amino acid substitutions that occur within a molecule during the course of evolution.

Although the neutral theory was at first thought to be in conflict with Darwinian natural selection, there is no necessary contradiction. Indeed, in *The Origin of Species*, Darwin (1872) comments:

> I am inclined to suspect that we see, at least in some [cases], variations which are of no service to the species, and which consequently have not been seized on and rendered definite by natural selection. . . . Variations neither useful nor injurious would not be affected by natural selection, and would be left either a fluctuating element, as perhaps we see in certain polymorphic species, or would ultimately become fixed. . . . We may easily err in attributing importance to characters, and in believing that they have been developed through natural selection; . . . many structures are now of no direct use to their possessors,

and may never have been of any use to their progenitors. . . . [On the other hand,] we are much too ignorant in regard to the whole economy of any organic being to say what slight modifications would be of importance or not.

The neutral theory has been of considerable importance in population genetics in stimulating the collection and analysis of data in attempts to evaluate its adequacy. Mathematical investigations of its implications have resulted in what can only be regarded as one of the most complete and elegant theories in all of biology. The next section summarizes some of the theoretical implications of the neutral theory and some of the data bearing on it. Although the neutral theory of molecular evolution has gained many adherents throughout the years, it is by no means universally accepted among population geneticists, and there are doubts and inconsistencies that remain to be resolved. Definition of the adequacy and range of applicability of the neutral theory remains an important objective in both theoretical and experimental population genetics.

Theoretical Principles of the Neutral Theory

The neutral theory is more precisely called the neutral mutation and random drift hypothesis because it concerns mutations that are so nearly selectively neutral in their effects that their fate is determined largely through random genetic drift. The most common theoretical model of random genetic drift has been described in Chapter 2. It assumes an ensemble of populations, each of finite size N, with nonoverlapping generations. Reproduction occurs through the N adult individuals producing an infinite pool of gametes, from which $2N$ are chosen at random to create the N zygotes of the next generation (see Figure 2 in Chapter 2). Some of the principal theoretical implications of such a population structure are summarized in the following list (Kimura 1983). Principles numbered 1, 2, 5, and 6 were demonstrated and discussed in Chapter 2 and are included here for completeness.

1. *If a population contains a neutral allele with allele frequency* p_0, *then the probability that the allele eventually becomes fixed equals* p_0. In particular, a newly arising neutral mutation occurs in just one copy, so the initial allele frequency is $p_0 = 1/(2N)$, and the probability of eventual fixation of the mutation is therefore $1/(2N)$.
2. *The steady-state rate at which neutral mutations are fixed in a population equals* μ, *where* μ *is the rate of occurrence of neutral mutations per gene per generation.* The rate μ is called the **neutral mutation rate**. It is noteworthy that the equilibrium rate of fixation does not involve the population size N. The reason is that the N cancels out — the overall rate is determined by the

product of the probability of fixation of new neutral mutations [1/(2N)] and the average number of new neutral mutations in each generation (2$N\mu$); hence [1/(2N)] × (2$N\mu$) = μ.

As you will see later in this chapter, the ultimate fate of mutations with small effects on fitness is also determined largely by random genetic drift, provided that the selection coefficients are small relative to the reciprocal of the effective population number. (The effective population number, defined in Chapter 2, corresponds roughly to the number of breeding individuals in the actual population; typically, the effective size of a population is smaller, and sometimes much smaller, than its actual size.) Since the relative importance of selection and random genetic drift is determined by the relative magnitudes of the selection coefficient and the effective population number, the rate of neutral mutation μ is itself affected by the effective population number. Perhaps surprisingly, the rate of substitution of neutral alleles is not affected by linkage to selected genes (Birky and Walsh 1988). Alleles with selection coefficients that are small in relation to effective population number (i.e., $|s| << 1/2N_e$) are nearly neutral, and their ultimate fate is determined largely by random genetic drift. Alleles with equally small selection coefficients occurring in much larger populations may be under the almost exclusive control of selection.

3. *The average time that occurs between consecutive neutral substitutions equals* $1/\mu$. This principle follows directly from the one above. If the steady-state rate of fixation is μ per unit time, the average length of time between substitutions will be the reciprocal, or $1/\mu$. By way of analogy, if a Swiss clock cuckoos 24 times per day, then the average length of time between cuckoos is 1/24th of a day, or one hour. However, the analogy with gene substitutions is improved if we assume a novelty clock in which the time interval between successive cuckoos is random (specifically, Poisson distributed), with a one-hour average.
4. *Analysis of the diffusion equation has shown that among newly arising neutral alleles that are destined to be fixed, the average time to fixation is* $4N_e$ *generations; among newly arising neutral alleles destined to be lost, the average time to loss is* $(2N_e/N)ln(2N)$ *generations.* In these equations, N_e and N refer, respectively, to the effective population number and the actual population number. The average times required for fixation or loss apply to newly arising alleles, which are necessarily present in just one copy, so $p_0 = 1/(2N)$. (The symbol "ln" in the second formula is the natural logarithm, which uses the base e.)

The implication of these formulas is that, on average, neutral mutations that are going to be fixed require a very long time for fixation to occur, but mutations destined to be lost are lost quite rapidly. This

contrast is apparent in Table 1. For a population of 1000 individuals, for example, with $N_e = 0.8N$, the average time to loss is 12.2 generations, whereas the average time to fixation is 3200 generations. Table 1 also illustrates that times to loss are not markedly affected by population size. Specifically, as the population size increases from 50 to 50,000 (a 1000-fold increase), the average times to loss increase from 7.4 generations to 18.4 generations (a 2.5-fold increase).

5. *In the absence of new mutations, random genetic drift decreases the average proportion of heterozygotes in an ensemble of populations at the rate of* $1/(2N_e)$ *per generation.* Eventually, of course, each subpopulation becomes fixed for one allele or another, and at this time the average heterozygosity across the entire ensemble is zero. Nevertheless, if mating occurs at random within each subpopulation, then the genotype frequencies within each subpopulation at any time are given approximately by the Hardy-Weinberg principle.
6. *If each neutral mutation creates an allele that is different from all others existing in the population in which it occurs, then at equilibrium, when the average number of new alleles gained through mutation is exactly offset by the average number lost through random genetic drift, the expected homozygosity equals* $1/(4N_e\mu + 1)$*, where* μ *is the neutral mutation rate.* The model of mutation in which each new allele is novel is the **infinite alleles model** of mutation. The quantity $4N_e\mu$, which shows up frequently in the neutral theory, is often denoted as θ. The equilibrium average homozygosity is therefore $1/(1 + \theta)$. Since the heterozygosity equals one minus the homozygosity,

Table 1. Average time to fixation or loss of neutral alleles.

N[a]	Average generations until loss[b]	Average generations until fixation[b]
50	7.4	160
100	8.5	320
250	9.9	800
500	11.1	1,600
1,000	12.2	3,200
5,000	14.7	16,000
10,000	15.8	32,000
50,000	18.4	160,000

[a] Tabulated values assume that $N_e = 0.8N$.
[b] Averages are for selectively neutral alleles destined to be lost or fixed, respectively.

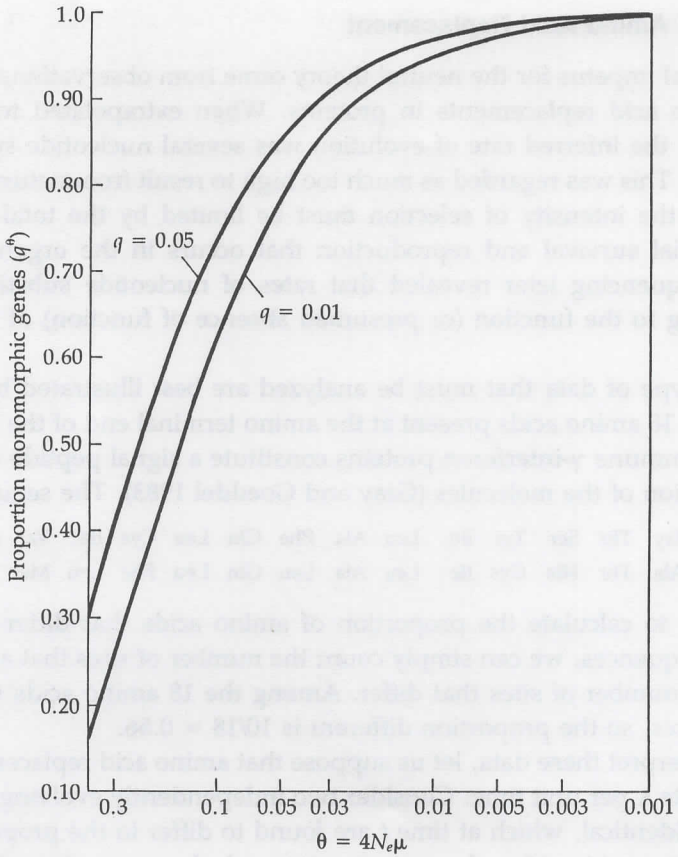


Figure 9. Approximate proportion of genes expected to be monomorphic assuming neutral mutation rate μ. A gene is considered monomorphic if the most common allele has a frequency greater than $1 - q$, where usually $q = 0.01$ or 0.05. Levels of electrophoretically detected protein polymorphism are usually consistent with values of θ in the range 0.03–0.3.

the average heterozygosity at equilibrium in the infinite alleles model equals $\theta/(1 + \theta)$.

7. *If monomorphism is defined as a gene for which the most common allele has a frequency greater than or equal to* $1 - q$, *where* q *is some small number such as 0.01 or 0.05, then at equilibrium the probability that a gene will be monomorphic is approximately* q^θ, *where* $\theta = 4N_e\mu$. Values of q^θ are plotted for $q = 0.01$ and $q = 0.05$ in Figure 9. Note that the scale of θ is logarithmic. In most organisms, the proportion of monomorphic genes discovered in electrophoretic surveys is consistent with values of θ in the range 0.03 to 0.30.

Rates of Amino Acid Replacement

The initial impetus for the neutral theory came from observations on the rate of amino acid replacements in proteins. When extrapolated to the entire genome, the inferred rate of evolution was several nucleotide substitutions per year. This was regarded as much too high to result from natural selection, because the intensity of selection must be limited by the total amount of differential survival and reproduction that occurs in the organism. Direct DNA sequencing later revealed that rates of nucleotide substitution vary according to the function (or presumed absence of function) of the nucleotides.

The type of data that must be analyzed are best illustrated by example. The first 18 amino acids present at the amino terminal end of the human and mouse immune γ-interferon proteins constitute a signal peptide that is used in secretion of the molecules (Gray and Goeddel 1983). The sequences are

Human:	Met	Lys	Try	Thr	Ser	Tyr	Ile	Leu	Ala	Phe	Gln	Leu	Cys	Ile	Val	Leu	Gly	Ser
Mouse:	Met	Asn	Ala	Thr	His	Cys	Ile	Leu	Ala	Leu	Gln	Leu	Phe	Leu	Met	Ala	Val	Ser

In order to calculate the proportion of amino acids that differ in the two signal sequences, we can simply count the number of sites that are the same and the number of sites that differ. Among the 18 amino acids there are 10 differences, so the proportion different is $10/18 = 0.56$.

To interpret these data, let us suppose that amino acid replacements occur at the rate λ per unit time. Consider two independently evolving sequences, initially identical, which at time t are found to differ in the proportion D_t of their amino acids. After the next time interval, the proportion of differences D_{t+1} is given by

$$D_{t+1} = (1 - D_t)(2\lambda) + D_t \qquad (7.14)$$

In this equation, $(1 - D_t)(2\lambda)$ is the proportion of sites, previously identical, in which one or the other underwent an amino acid replacement during the time interval in question, which must be added to the already existing differences D_t in order to give the total. (The equation ignores the unlikely possibility of an amino acid replacement making two previously different amino acid sites identical.) The factor of 2 is present because the total time for evolution is $2t$ units (t units in each lineage after the split), which is illustrated in Figure 10.

Equation 7.14 suggests the differential equation

$$dD/dt = D_{t+1} - D_t = 2\lambda - 2\lambda D(t) \qquad (7.15)$$

which has the solution

$$D_t = 1 - e^{-2\lambda t} \qquad (7.16)$$

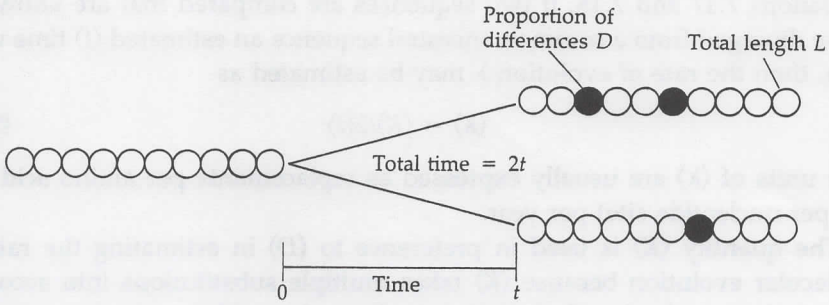


Figure 10. Two amino acid or nucleotide sequences which have each undergone independent evolution from a common ancestor for t time units are separated by a total time of $2t$ units, because there are t units in each lineage after the split. The proportion of sites which differ in the sequence is denoted $\langle D \rangle$ and the total number of sites L. In this particular example, $L = 10$ and $\langle D \rangle = 3/10$.

An alternative argument can be used to derive Equation 7.16 without resorting to differential equations. If λ is the rate of amino acid replacement per unit time, then the probability that a particular site remains unsubstituted for t consecutive intervals along each of two independent lineages is $(1 - \lambda)^{2t}$, which is approximately equal to $e^{-2\lambda t}$, provided that λt is not too large. Thus, the probability D_t of one or more replacements occurring in t units of time after divergence is approximately $1 - e^{-2\lambda t}$, which is Equation 7.16.

Since λ is the rate of amino acid replacement per unit time, the expected proportion of differences between two sequences at any time t is

$$K = 2\lambda t \qquad \textbf{(7.17)}$$

where the factor of 2 is again present because the total time for evolution is $2t$ units (Figure 10).

Substituting K from Equation 7.17 into Equation 7.16 and rearranging yields the following estimate $\langle K \rangle$ of K,

$$\langle K \rangle = -\ln(1 - \langle D \rangle) \qquad \textbf{(7.18)}$$

where $\langle D \rangle$ is the observed proportion of sites in which two sequences differ. If the sequences under comparison are L amino acids in length, then the estimated variance $\langle \mathrm{Var}(\langle K \rangle) \rangle$ of $\langle K \rangle$ is approximately

$$\langle \mathrm{Var}(\langle K \rangle) \rangle = \langle D \rangle / [(1 - \langle D \rangle) L] \qquad \textbf{(7.19)}$$

The rate of evolution at the molecular level is given by the amount of sequence divergence that occurs per unit of time. Thus, as suggested by

Equations 7.17 and 7.18, if two sequences are compared that are known to have diverged from a common ancestral sequence an estimated $\langle t \rangle$ time units ago, then the rate of evolution λ may be estimated as

$$\langle \lambda \rangle = \langle K \rangle / 2\langle t \rangle \tag{7.20}$$

The units of $\langle \lambda \rangle$ are usually expressed as replacements per amino acid site (or per nucleotide site) per year.

The quantity $\langle K \rangle$ is used in preference to $\langle D \rangle$ in estimating the rate of molecular evolution because $\langle K \rangle$ takes multiple substitutions into account. Over long periods of evolutionary time, the amino acid present at a particular site may be replaced several times, first by one alternative, then another, then still another, and perhaps, at some stage, even return to the amino acid originally present at the site. When comparing two sequences, only the sites that are different can be identified. Sites that are identical at the present time may include some that were different in the past, and sites that are different at the present time might have undergone more than one substitution. The quantity $\langle D \rangle$ is determined only by the proportion of differences between the sequences observed at the present time. The estimate $\langle K \rangle$ makes a correction for multiple substitutions, but at the cost of introducing assumptions that the substitutions occur independently and at the same rate through time. For relatively short intervals of evolutionary time, during which multiple substitutions remain uncommon, the correction is minor, and the value of $\langle K \rangle$ is close to that of $\langle D \rangle$. Over longer intervals, when multiple substitutions have occurred, the correction is important, and the assumptions on which it is based must be evaluated critically. Correction for multiple substitution events is even more important for nucleotides than it is for amino acids. With amino acids, the probability of a random replacement returning an amino acid site to its original identity is 1/20 (assuming uniform frequencies), whereas for nucleotides it is 1/4.

Application of Equations 7.17 and 7.18 may be illustrated using the data in the preceding example to estimate the average rate of amino acid replacement in the signal peptide of γ-interferon during the divergence of mice and humans. Based on fossil evidence, the separation of these species occurred approximately 80 million years ago. For the signal peptide, $\langle D \rangle = 0.56$ and $\langle K \rangle = -\ln(1 - 0.56) = 0.82$. The estimated rate of evolution is therefore, $\langle \lambda \rangle = 0.82/[2 \times (80 \times 10^6)] = 5.1 \times 10^{-9}$ amino acid replacements per amino acid site per year. The standard deviation of $\langle K \rangle$ is estimated as equal to $[0.56/(0.44 \times 18)]^{1/2} = 0.27$. With such a small sample size, the estimates could not ordinarily be taken too literally. However, in this case, the average rate for the signal sequence is very close to the average rate for the molecule as a whole. For γ-interferon, among 155 amino acid sites there are 91 differences, giving $\langle K \rangle = 0.88 \pm 0.22$ and an average rate of $\langle \lambda \rangle = 5.5 \times 10^{-9}$ amino acid replacements per amino acid site per year.

Rates of amino acid replacement vary over a 500-fold range in different proteins. The rate of amino acid replacement in γ-interferon is one of the fastest rates known (Li et al. 1985a). Among the slowest rates is that of histone H4, for which $\langle\lambda\rangle = 0.01 \times 10^{-9}$ per year. The average rate among a large number of proteins is very close to the rate found in hemoglobin, which is approximately 1×10^{-9} amino acid replacements per amino acid site per year.

To be concrete about the interpretation of the rate of amino acid replacement, consider a protein exactly 100 amino acids in length in which the rate of amino acid replacement per amino acid site equals 1.0×10^{-9} per year. For the entire protein, the rate of replacement equals $100 \times 1.0 \times 10^{-9} = 1 \times 10^{-7}$ per year. In two different species, therefore, the protein would accumulate amino acid differences at the rate of one replacement every 5 million years since their divergence from a common ancestor [because $(5 \times 10^6)(2)(1 \times 10^{-7}) = 1.0$]. More generally, the expected proportion of amino acid differences between the proteins at a time t years after their divergence may be calculated from Equation 7.16 with $\lambda = 1.0 \times 10^{-9}$.

When protein sequences are compared between species and estimates made using Equations 7.18 and 7.19, it is usually assumed that the sequences being compared are truly representative of the species. However, as emphasized in Chapter 1, genetic variation occurs in virtually all natural populations, and the protein or gene being studied may be polymorphic. Estimation of rates of evolution become much more difficult when intraspecific genetic variation is taken into account. Indeed, intraspecific genetic variation is usually ignored in the knowledge, or sometimes only the hope, that the magnitude of intraspecific variation is small in comparison with that between the species being compared.

Rates of Nucleotide Substitution

Nucleotide sequences are analyzed in the same manner as amino acid sequences, but the equation analogous to Equation 7.14 is slightly more complicated because it has to correct for cases in which a substitution makes two previously different nucleotide sites identical. The correction is significant for nucleotide sequences because an expected 1/3 of random substitutions will make two previously different nucleotides identical. The correction is usually unnecessary for proteins because only 1/19 of random replacements can make two previously different amino acids identical.

Various models of nucleotide substitution have been studied; they differ primarily in the assumptions about rates of mutation between pairs of nucleotides. The simplest model is one in which mutation occurs at a rate λ per unit time, and each nucleotide is equally likely to mutate to any other (Jukes and Cantor 1969). In this case the equation analogous to Equation 7.18 is

$$\langle k \rangle = -(3/4)\ln(1 - 4\langle D \rangle/3) \qquad (7.21)$$

where k is the expected proportion of nucleotide sites that differ between two sequences at a time t units after their evolutionary separation. By analogy with protein evolution (Figure 10), $\langle D \rangle$ is the observed proportion of L nucleotide sites in which the sequences differ. The variance $\mathrm{Var}(\langle k \rangle)$ of the estimate can be estimated as

$$\langle Var(\langle k \rangle) \rangle = \frac{\langle D \rangle(1 - \langle D \rangle)}{L(1 - 4\langle D \rangle/3)^2} \qquad (7.22)$$

Application of Equations 7.21 and 7.22 to nucleotide sequences is illustrated in the following example. The coding region of the *trpA* genes in strains of the related enteric bacteria *Escherichia coli* strain K12 and *Salmonella typhimurium* strain LT2 were sequenced and compared (Nichols and Yanofsky 1979). The *trpA* gene codes for one of the subunits of the enzyme tryptophan synthetase used in the synthesis of tryptophan. We wish to estimate the amount of nucleotide divergence $\langle k \rangle$ and amino acid divergence $\langle K \rangle$ and their standard deviations. In the sequences, nucleotides that differ are marked with asterisks and amino acids that differ are underlined.

	*	*	*								*		
K12:	GTC	GCA	CCT	ATC	TTC	ATC	TGC	CCG	CCA	AAT	GCC	GAT	GAC
	Val	Ala	Pro	Ile	Phe	Ile	Cys	Pro	Pro	Asn	Ala	Asp	Asp
LT2:	ATC	GCG	CCG	ATC	TTC	ATC	TGC	CCG	CCA	AAT	GCG	GAT	GAC
	Ile	Ala	Pro	Ile	Phe	Ile	Cys	Pro	Pro	Asn	Ala	Asp	Asp

	*	*			* *	*	
K12:	GAC	CTG	CTG	CGC	CAG	ATA	GCC
	Asp	Leu	Leu	Arg	Gln	Ile	Ala
LT2:	GAT	CTT	CTG	CGC	CAG	GTC	GCA
	Asp	Leu	Leu	Arg	Gln	Val	Ala

For the amino acid sequences, $L = 20$ and $\langle D \rangle = 2/20 = 0.10$; thus $\langle K \rangle = -\ln(0.90) = 0.105$ with standard deviation 0.074. For the nucleotide sequences, $L = 60$ and $\langle D \rangle = 9/60 = 0.15$; thus $\langle k \rangle = -(3/4)\ln(0.8) = 0.167$ with standard deviation 0.058. Assuming that *Escherichia* and *Salmonella* diverged at around the time of the mammalian radiation 80 million years ago, the rates of evolution are $0.167/(2 \times 80 \times 10^6) = 1.04 \times 10^{-9}$ nucleotide substitutions per year and $0.105/(2 \times 80 \times 10^6) = 0.66 \times 10^{-9}$ amino acid replacements per year. In the gene as a whole, the values are $\langle k \rangle = 0.300$ for nucleotide substitutions and $\langle K \rangle = 0.162$ for amino acid replacements.

MOLECULAR CLOCKS

Although the rate of nucleotide and amino acid substitution varies among different genes, the average rate of molecular evolution is remarkably uni-

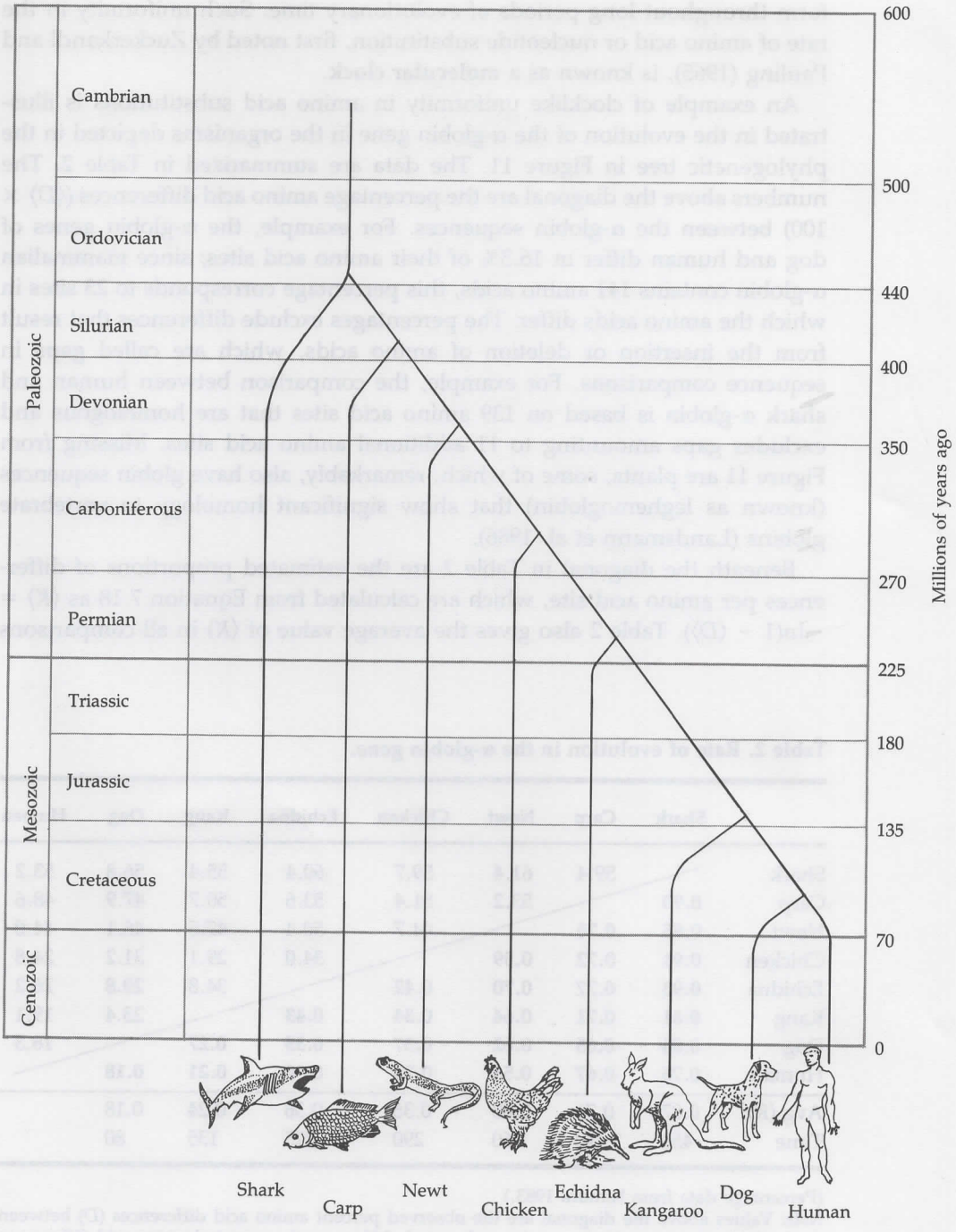


Figure 11. Phylogenetic relationships among eight vertebrate species and their approximate times of evolutionary divergence. (From Kimura 1983.)

form throughout long periods of evolutionary time. Such uniformity in the rate of amino acid or nucleotide substitution, first noted by Zuckerkandl and Pauling (1965), is known as a **molecular clock**.

An example of clocklike uniformity in amino acid substitutions is illustrated in the evolution of the α-globin gene in the organisms depicted in the phylogenetic tree in Figure 11. The data are summarized in Table 2. The numbers above the diagonal are the percentage amino acid differences ($\langle D \rangle \times 100$) between the α-globin sequences. For example, the α-globin genes of dog and human differ in 16.3% of their amino acid sites; since mammalian α-globin contains 141 amino acids, this percentage corresponds to 23 sites in which the amino acids differ. The percentages exclude differences that result from the insertion or deletion of amino acids, which are called **gaps** in sequence comparisons. For example, the comparison between human and shark α-globin is based on 139 amino acid sites that are homologous and excludes gaps amounting to 11 additional amino acid sites. Missing from Figure 11 are plants, some of which, remarkably, also have globin sequences (known as leghemoglobin) that show significant homology to vertebrate globins (Landsmann et al. 1986).

Beneath the diagonal in Table 2 are the estimated proportions of differences per amino acid site, which are calculated from Equation 7.18 as $\langle K \rangle = -\ln(1 - \langle D \rangle)$. Table 2 also gives the average value of $\langle K \rangle$ in all comparisons

Table 2. Rate of evolution in the α-globin gene.

	Shark	Carp	Newt	Chicken	Echidna	Kang	Dog	Human
Shark		59.4	61.4	59.7	60.4	55.4	56.8	53.2
Carp	**0.90**		53.2	51.4	53.6	50.7	47.9	48.6
Newt	**0.95**	**0.76**		44.7	50.4	47.5	46.1	44.0
Chicken	**0.91**	**0.72**	**0.59**		34.0	29.1	31.2	24.8
Echidna	**0.93**	**0.77**	**0.70**	**0.42**		34.8	29.8	26.2
Kang	**0.81**	**0.71**	**0.64**	**0.34**	**0.43**		23.4	19.1
Dog	**0.84**	**0.65**	**0.62**	**0.37**	**0.35**	**0.27**		16.3
Human	**0.76**	**0.67**	**0.58**	**0.28**	**0.30**	**0.21**	**0.18**	
Avg $\langle K \rangle$	0.87	0.71	0.63	0.35	0.36	0.24	0.18	
Time	450	410	360	290	225	135	80	

(Percentage data from Kimura 1983.)
Note: Values above the diagonal are the observed percent amino acid differences (D) between the α-globin sequences in the species, values in **boldface** are the expected amino acid differences per site [$\langle K \rangle = -\ln(1 - D)$]. Average values of $\langle K \rangle$ and the estimated times of divergence (in millions of years) are given at the bottom of the table. Abbreviation: Kang, kangaroo.

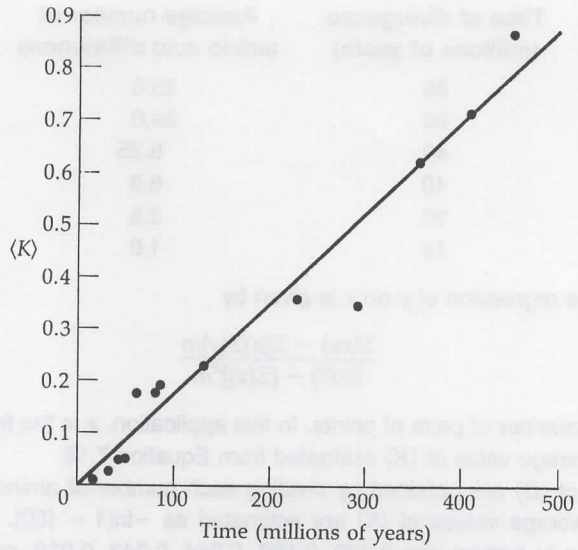


Figure 12. Relation between estimated number of amino acid substitutions in α-globin $\langle K \rangle$ between pairs of the vertebrate species in Figure 11, against time since each pair diverged from a common ancestor. The straight line is expected based on a uniform rate of amino acid substitution during the entire period. (From Kimura 1983.)

with the shark, carp, newt, chicken, echidna, kangaroo, and dog, respectively, and the divergence times from the bifurcations in Figure 11.

The average proportion of differences per site is plotted against divergence time in Figure 12. The very close fit to a straight line is evident. Since the divergence time is exactly half of the total time available for evolution (Figure 10), the rate of evolution λ can be estimated as one-half times the slope of the line in Figure 12. For these data, the slope is 1.8×10^{-9}, and therefore $\langle \lambda \rangle = 0.9 \times 10^{-9}$ amino acid substitutions per amino acid site per year. The good fit of the points to the straight line indicates that the actual rate of α-globin evolution has deviated little from the average for the past 450 million years.

Calculating the Average Rate of Molecular Evolution

E

The β-globin molecule in primates contains 146 amino acids, and estimates of the number of amino acid differences among various primates are tabulated below (data from Kimura 1983). In order to calculate the average rate of evolution of the β-globin molecule in primates, linear regression can be applied to these data.

Time of divergence (millions of years)	Average number of amino acid differences
85	25.5
60	24.0
42	6.25
40	6.0
30	2.5
15	1.0

The slope of the regression of y on x is given by

$$\frac{\Sigma(xy) - \Sigma(x)\Sigma(y)/n}{\Sigma(x^2) - [\Sigma(x)]^2/n}$$

where n is the number of pairs of points. In this application, x is the time of divergence and y is the average value of $\langle K \rangle$ estimated from Equation 7.18.

The values of $\langle D \rangle$ are obtained by dividing each number of amino acid differences by 146, and average values of $\langle K \rangle$ are estimated as $-\ln(1 - \langle D \rangle)$. The average $\langle K \rangle$ values, from top to bottom, are 0.192, 0.180, 0.044, 0.042, 0.018, and 0.007, respectively. These are the y values in the linear regression; the x values are the divergence times, as in Figure 12. Altogether there are $n = 6$ points. In this case, $\Sigma(xy) = 3.126 \times 10^7$, $\Sigma(x) = 2.720 \times 10^8$, $\Sigma(y) = 0.483$, and $\Sigma(x^2) = 1.531 \times 10^{16}$. The slope of the regression is 3.16×10^{-9}, and the rate of evolution is half of this, or 1.58×10^{-9} amino acid replacements per amino acid site per year. This estimate is reasonably close to the value of 0.9×10^{-9} per year calculated for α-globin. (Note: Rather than calculate $\langle K \rangle$ from the average number of amino acid differences, it would be more accurate to calculate $\langle K \rangle$ for each species comparison and then take the average; however, in this example, it makes very little difference.)

Although the rate of evolution is approximately constant for a given gene, the rates vary among genes. Genes whose function is well buffered from the environment generally have a slower rate of substitution than genes whose products have a premium on variability. The extremes are represented by histone H4 at the low end, and γ-interferon at the high end, with globin genes near the middle of the spectrum. In short, the molecular clocks for different genes "tick" at different rates. The rate of the molecular clock also varies among taxonomic groups (Britten 1986). For example, the insulin gene evolved much more rapidly in the evolutionary line leading to the guinea pig than in other evolutionary lines (King and Jukes 1969), and the C-type viral sequences integrated into the primate genome evolved at twice the rate in Asian primates as in African primates (Benveniste 1985).

Furthermore, molecular evolutionary clocks are "sloppy" clocks in that they demonstrate an enhanced variance in the evolutionary rate. In a theoretically perfect molecular clock driven by a random process identical to that

of radioactive decay (a Poisson process), the variance in the rate of ticking would be equal to the average rate of ticking. However, in molecular evolutionary clocks, the estimated variance is typically about twice the magnitude of the mean. Gillespie's (1986a,b) analysis suggests that the ratio of the variance to the mean might be even greater (see Box F).

The possibility that gene substitutions might occur at an approximately constant rate gave some credence to the simplest version of the neutral theory. Theoretical principle number 2 discussed earlier in this chapter states that the expected rate of substitution of neutral alleles equals the rate of mutation μ to neutral alleles. Therefore, on the face of it, the occurrence of molecular clocks would seem to support the neutral theory. However, one observed feature of molecular evolutionary clocks is that their rate is approximately constant in a time scale measured in *years*. This is quite unexpected because mutation rates are thought to be more nearly constant when measured in *generations*. However, the appropriate time scale of molecular evolution is not completely settled (Easteal 1985), as there is some evidence that the rate of synonymous substitution in genes in the rodent lineage (short generation time) might be about two times as rapid as occurs in the same genes in the human lineage (Wu and Li 1985; Li and Wu 1987). Evidence from immunoglobulin genes further suggests that among mammals, the primate lineage has the slowest rate of nucleotide substitution (Sakoyama et al. 1987).

Comparisons among nucleotide substitution rates of nuclear DNA, chloroplast DNA (cpDNA), and mitochondrial DNA (mtDNA) in plants reveal clear differences, with mtDNA showing less than one-third the substitution rate of cpDNA, which in turn has about half the substitution rate of nuclear genes (Wolfe et al. 1987). Rates of nucleotide substitution in gene families can also vary widely (Harris et al. 1986).

Even if true, a nearly constant rate of gene substitution per year is not necessarily in conflict with a constant rate of neutral mutation per generation. The reason is that organisms with short generation times tend to be small and to maintain large population sizes. In such organisms, the proportion of nearly neutral mutations will be reduced because effective neutrality requires that $Ns << 1$, where s is the selection coefficient against the mutation. However, the smaller proportion of nearly neutral mutations in these organisms is offset against the occurrence of more mutations per unit time than in larger organisms, because the generation time is shorter. Thus, the effects of short generation time and larger population size act in opposite directions and tend to cancel each other out (Crow 1985).

It has been argued that the apparent clocklike regularity of gene substitutions is actually an artifact of averaging evolutionary rates over long periods of time. A continuous record of gene substitutions might indicate that sub-

stitutions occur in clusters rather than at nearly regular intervals. This kind of pattern is suggested by the analyses of Gingerich (1986) and Gillespie (1986a,b).

F Is the Molecular Clock Episodic?

The neutral theory predicts that nucleotide substitutions should occur according to a Poisson process with an average rate equal to the neutral mutation rate. In a Poisson process, the mean and the variance in the rate of substitution are equal. Tests based on equality of mean and variance by Ohta and Kimura (1971) and Langley and Fitch (1974) revealed that the variance in substitution rate was slightly greater than the mean. More recently, Gillespie (1986a) examined the ratio R of the variance to the mean in a set of four nuclear and five mitochondrial genes in mammals, and found that R ranged from 0.16 to 35.55. (The value of 35.55 is for cytochrome oxidase II, which shows 65 amino acid differences between human and mouse, 61 differences between human and cow, and only 21 differences between mouse and cow.) Gillespie argued that the large range of R implies a sixfold difference within mammalian lineages in rates of nucleotide substitution. This fluctuation in substitution rate has been called an "episodic clock," characterized by periods of stasis alternating with periods of rapid substitution (Gillespie 1984). Models that incorporate natural selection in a changing environment can be made to fit these data (Gillespie 1986b). Fluctuating selection cannot explain the observation that synonymous sites have values of R that are comparable to those of nonsynonymous sites, and significantly exceed 1. Takahata (1987) has developed three modifications of the neutral model that show inflated variance. The first modification assumes that multiple substitutions are not strictly independent of each other (that is, they occur like a compound Poisson process). The second modification assumes a combination of deleterious mutations and population bottlenecks. Finally, in a third modification (called the "fluctuating neutral space" model), changes in selective constraints among lineages result in variation in substitution rate among lineages. Whether high values of R are grounds for rejecting the neutral theory is disputed (Gillespie 1988; Takahata 1988). Resolution may come with more analysis of sequence data and a deeper look into neutral models that have an excess variance in the substitution rate. In the meantime, the observation of an excess variance-to-mean ratio for substitution rates is a nagging problem not fully explained by the theory.

The explanation of molecular clocks as an artifact of averaging has been examined with data on the occurrence of gene substitutions in short-lived organisms. An example occurs in the *NS* genes of human influenza A virus. Influenza is an RNA-containing virus, and the *NS* genes code for nonstructural proteins that are not involved in immune recognition by the host. Strains of influenza have been isolated over the past 50 years, and, except for the earliest isolates, have been maintained in refrigeration.

Figure 13 shows the number of nucleotide differences observed in the

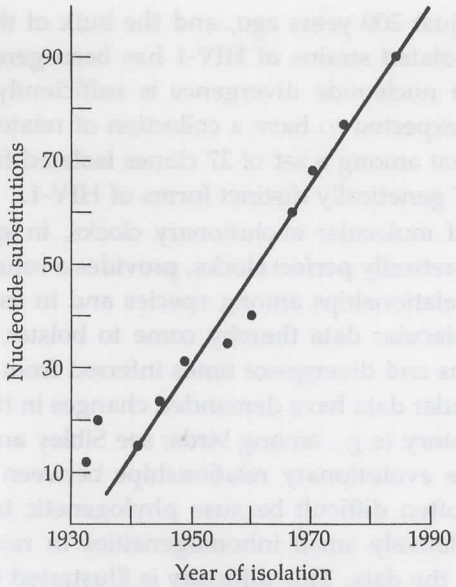


Figure 13. Molecular evolution in the *NS* genes of influenza virus determined from strains isolated and stored during the past 60 years. The total rate of evolution in the 890-nucleotide sequence averages 1.73 ± 0.08 nucleotide substitutions per year, and the rate is remarkably uniform. (From Buonagurio et al. 1986.)

influenza *NS* genes plotted against the year of isolation of the virus containing them. The rate of gene substitution averages $\langle\lambda\rangle = 1.94 \pm 0.09 \times 10^{-3}$ nucleotide substitutions per nucleotide site per year. Although the rate of gene substitution is about 10^6-fold faster than observed in germline genes in eukaryotes, it is nevertheless approximately constant during the period available for study. The extraordinary rate of evolution in influenza virus is thought to be related to a high rate of spontaneous mutation resulting from errors in replication (Holland et al. 1982). As in many other RNA-based viruses, the RNA replicase enzyme that replicates the influenza genome lacks a proofreading function. The rate of mutation is estimated to be approximately 0.01 per nucleotide site per year, but the nucleotide substitution rate is evidently much lower than this due to purifying selection (Saitou and Nei 1986). Another interesting feature of influenza virus molecular evolution is the observation of strains that seem to have been "frozen" in their nucleotide sequence for a period of years (Nakajima et al. 1978; Hayashida et al. 1985).

Rapid rates of gene substitution can be of immense medical significance. Yokoyama et al. (1988) estimated the rate of substitution at the *pol* gene of the human immunodeficiency virus that causes AIDS as 0.5×10^{-3} per nucleotide site per year. The time of divergence between HIV-1 and HIV-2

was estimated at just 200 years ago, and the bulk of the genetic variability among recently isolated strains of HIV-1 has been generated in the last 20 years. The rate of nucleotide divergence is sufficiently rapid that infected people would be expected to have a collection of related clones. Saag et al. (1988) observed that among a set of 27 clones isolated from a single individual, there were 17 genetically distinct forms of HIV-1.

The existence of molecular evolutionary clocks, in spite of their greater variation than theoretically perfect clocks, provides a valuable tool in inferring the phylogenetic relationships among species and in estimating their times of divergence. Molecular data thereby come to bolster, and in some cases refine, relationships and divergence times inferred from the fossil record. In a few cases, molecular data have demanded changes in the established views of evolutionary history (e.g., among birds; see Sibley and Ahlquist 1986).

Determining the evolutionary relationships between organisms that are closely related is often difficult because phylogenetic inferences can be affected by even relatively small inhomogeneities in rates of divergence or sampling errors in the data. This difficulty is illustrated in the case of higher primates. Figure 14 shows the two most popular phylogenetic hypotheses for the relationships among humans, chimpanzees, and gorillas. Phylogeny (a) groups humans and chimpanzees; phylogeny (b) groups chimpanzees and gorillas.

Based on restriction fragment length differences in mtDNA, Templeton (1983) rejected hypothesis (a) in favor of (b). However, the validity of the statistical analysis was contested by Nei and Tajima (1985), who concluded that neither phylogeny could be rejected based on the mtDNA data. On the other side, analysis of differences in nuclear DNA sequences, as judged by their ability to form thermally stable hybrid molecules, led Sibley and Ahlquist (1984) to reject phylogeny (b). These data were reanalyzed by Templeton (1985), who concluded that rejection of (b) was unwarranted. His statistical methods were challenged by Saitou (1986) and Ruvolo and Smith (1986), and defended by Templeton (1986). More recent data by Sibley and Ahlquist (1987) also support phylogeny (a). Discordant phylogenies have also been constructed from nuclear and mitochondrial DNA of Hawaiian *Drosophila* (DeSalle and Giddings 1986).

Holmquist et al. (1988) used the methods of Lake (1987) to combine mtDNA and globin sequence data, and found that (a) was favored, although a trifurcation could not be rejected. The situation is therefore very much in dispute, but identifying the correct phylogeny is important in determining whether human traits such as the bipedal gait and human limb proportions should be considered as evolutionarily ancestral or derived (Templeton 1983). The essence of the problem is that the species are so closely related that neither the presently available data nor the methods of analysis can make

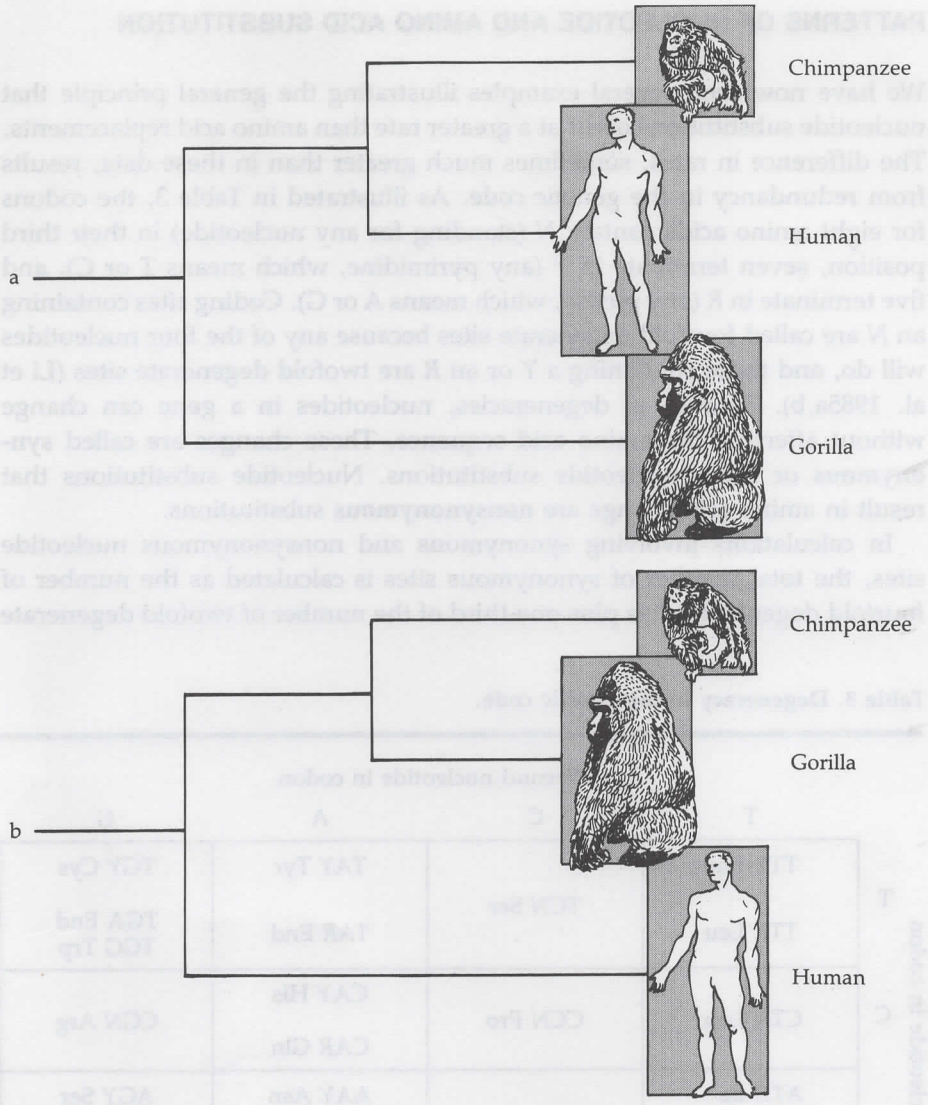


Figure 14. Two popular models for the phylogenetic relationship among humans, chimpanzees, and gorillas. Neither model can be convincingly rejected based on current data.

the critical distinctions. The problem is made even more difficult by the likelihood that there are polymorphisms shared by the species, implying that the polymorphisms predated the speciation. This is not an unprecedented possibility, since polymorphisms at the major histocompatibility loci are shared across species (McConnel et al. 1988).

We have now seen several examples illustrating the general principle that nucleotide substitutions occur at a greater rate than amino acid replacements. The difference in rates, sometimes much greater than in these data, results from redundancy in the genetic code. As illustrated in Table 3, the codons for eight amino acids contain *N* (standing for any nucleotide) in their third position, seven terminate in *Y* (any pyrimidine, which means T or C), and five terminate in *R* (any purine, which means A or G). Coding sites containing an *N* are called fourfold degenerate sites because any of the four nucleotides will do, and those containing a *Y* or an *R* are twofold degenerate sites (Li et al. 1985a,b). Because of degeneracies, nucleotides in a gene can change without affecting the amino acid sequence. These changes are called **synonymous** or silent nucleotide substitutions. Nucleotide substitutions that result in amino acid change are **nonsynonymous** substitutions.

In calculations involving synonymous and nonsynonymous nucleotide sites, the total number of synonymous sites is calculated as the number of fourfold degenerate sites plus one-third of the number of twofold degenerate

Table 3. Degeneracy in the genetic code.

First nucleotide in codon	Second nucleotide in codon: T	C	A	G
T	TTY Phe TTR Leu	TCN Ser	TAY Tyr TAR End	TGY Cys TGA End TGG Trp
C	CTN Leu	CCN Pro	CAY His CAR Gln	CGN Arg
A	ATH Ile ATG Met	ACN Thr	AAY Asn AAR Lys	AGY Ser AGR Arg
G	GTN Val	GCN Ala	GAY Asp GAR Glu	GGN Gly

Note: In this representation of the standard genetic code, the symbol *N* stands for any nucleotide (T, C, A, or G), the symbol *Y* for any pyrimidine (T or C), and the symbol *R* for any purine (A or G). The *H* in the set of codons for isoleucine (Ile) stands for "not-G" (T, C or A). Degeneracies are as follows: *N* represents a fourfold degenerate site, *Y* and *R* represent twofold degenerate sites. The *H* in the set of codons for isoleucine is considered as twofold degenerate, as are the first nucleotides in four leucine codons (TTA, TTG, CTA, and CTG) and four arginine codons (CGA, CGG, AGA, and AGG). All other nucleotides are nondegenerate.

sites. The total number of nonsynonymous sites in a coding region is defined as the number of nondegenerate sites (nucleotides in which any change results in an amino acid substitution), plus two-thirds of the number of twofold degenerate sites (the latter because, with random mutation at twofold degenerate sites, two-thirds of the mutations are expected to result in amino acid changes). These conventions are illustrated in Box G.

Paralleling the evolutionary rates for amino acid substitutions, the rates of nonsynonymous nucleotide substitution vary tremendously among different proteins. Among the slowest rates is that of histone H4, for which $\langle\lambda\rangle = 0.004 \times 10^{-9}$ substitutions per nonsynonymous nucleotide site per year, and among the fastest is that of γ-interferon, for which $\langle\lambda\rangle = 2.80 \times 10^{-9}$ substitutions per nonsynonymous nucleotide site per year. The average rate among a large number of proteins is very close to the rate found in hemoglobin, which is 0.87×10^{-9} substitutions per nonsynonymous nucleotide site per year (Figure 15). As in the examples given here, rates of nonsynonymous nucleotide substitution are usually quite similar to the rates of amino acid replacement in the same genes.

In contrast with the highly variable rates of nonsynonymous nucleotide substitutions among proteins, the rates of synonymous substitution are much more uniform. For example, in mammalian genes, the fastest rate of synonymous substitution is only 3–4 times greater than the slowest rate (Figure 15). However, the average rate, $\langle\lambda\rangle = 4.7 \times 10^{-9}$ substitutions per synonymous site per year, is not only greater than the average rate of nonsynonymous substitutions, but is greater than the fastest known rate of nonsynonymous substitutions (γ-interferon).

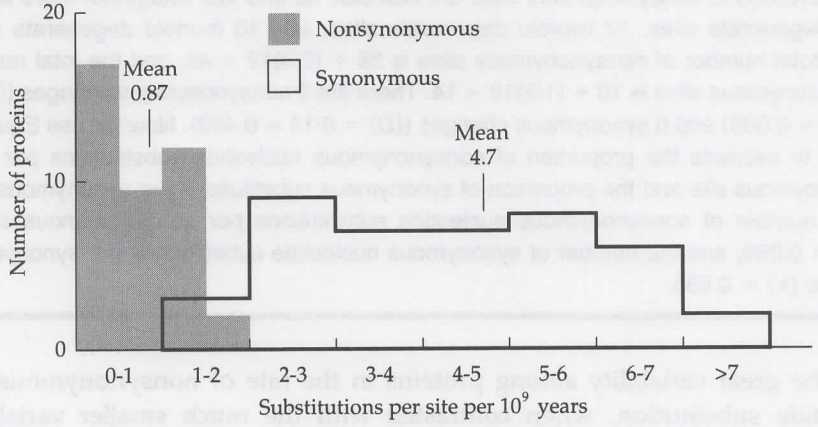


Figure 15. Distributions of rates of synonymous and nonsynonymous substitutions in protein-coding regions in animal genes. (Data from Li et al. 1985a.)

G Calculating the Number of Synonymous and Nonsynonymous Sites

Here we will use the degeneracy values in Table 3 to determine the degeneracy of each nucleotide in the region of the *Escherichia coli* strain K12 *trpA* gene given earlier. For each difference between *E. coli* and *Salmonella typhimurium* strain L2, the difference is synonymous if the site is fourfold degenerate or if it is twofold degenerate and the change is a transition (that is, A to G or the reverse, or T to C or the reverse). The difference is nonsynonymous if the site is nondegenerate or if it is twofold degenerate and the change is a transversion (that is, A or G to T or C). Equation 7.21 is used to estimate the proportion of nonsynonymous nucleotide substitutions per nonsynonymous site and the proportion of synonymous substitutions per synonymous site. The degeneracy assignments are therefore as follows:

	*	*	*								*		
	004	004	004	002	002	002	002	004	004	002	004	002	002
K12:	GTC	GCA	CCT	ATC	TTC	ATC	TGC	CCG	CCA	AAT	GCC	GAT	GAC
	Val	Ala	Pro	Ile	Phe	Ile	Cys	Pro	Pro	Asn	Ala	Asp	Asp
LT2:	ATC	GCG	CCG	ATC	TTC	ATC	TGC	CCG	CCA	AAT	GCG	GAT	GAC
	Ile												
	N	S	S								S		

	*	*			* *	*	
	002	204	204	004	002	002	004
K12:	GAC	CTG	CTG	CGC	CAG	ATA	GCC
	Asp	Leu	Leu	Arg	Gln	Ile	Ala
LT2:	GAT	CTT	CTG	CGC	CAG	GTC	GCA
						Val	
	S	S				N N	S

The stars above indicate differences with *Salmonella* and the letters below indicate which changes are nonsynonymous and which are synonymous. The two amino acid substitutions at nonsynonymous sites are indicated Ile and Val. Altogether there are 38 nondegenerate sites, 12 twofold degenerate sites, and 10 fourfold degenerate sites. The total number of nonsynonymous sites is $38 + (2/3)12 = 46$, and the total number of synonymous sites is $10 + (1/3)12 = 14$. There are 3 nonsynonymous changes ($\langle D \rangle = 3/46 = 0.065$) and 6 synonymous changes ($\langle D \rangle = 6/14 = 0.429$). Now we use Equation 7.21 to estimate the proportion of nonsynonymous nucleotide substitutions per nonsynonymous site and the proportion of synonymous substitutions per synonymous site. The number of nonsynonymous nucleotide substitutions per nonsynonymous site is $\langle \lambda \rangle = 0.068$, and the number of synonymous nucleotide substitutions per synonymous site is $\langle \lambda \rangle = 0.635$.

The great variability among proteins in the rate of nonsynonymous nucleotide substitution, when contrasted with the much smaller variability found in the rate of synonymous substitutions, is illustrated graphically in Figure 16. This disparity has been used as evidence in favor of the neutral

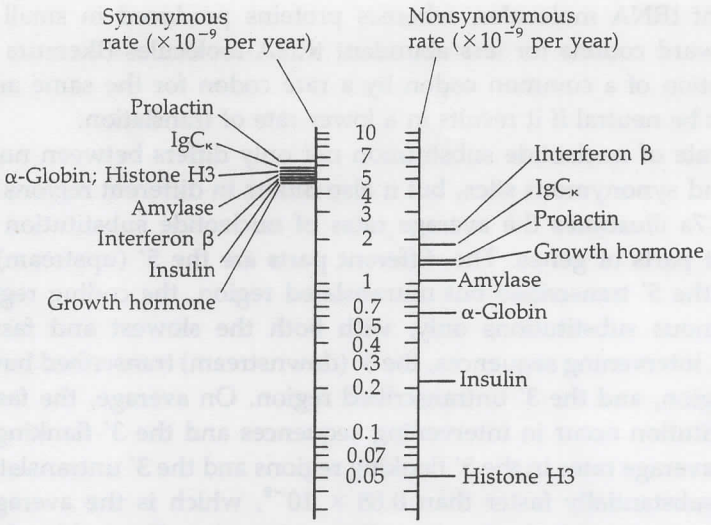


Figure 16. Comparison of rates of synonymous and nonsynonymous nucleotide substitutions. Synonymous rates are generally much faster and much more uniform than nonsynonymous rates. (From Kimura 1986.)

theory. Interpreted according to the neutral theory, the variation in rates occurs because there are selective constraints on amino acid substitutions that do not operate as strongly on synonymous nucleotide substitutions.

Not just any amino acid will serve at a particular position in a protein molecule, because each amino acid must participate in the chemical interactions that fold the molecule into its three-dimensional shape and give the molecule its specificity and ability to function. The need for proper chemical interactions and folding constrains the acceptable amino acids that can occupy each site. Although some amino acid substitutions may be functionally equivalent or nearly equivalent, many more are expected to impair protein function to such an extent that they reduce the fitness of the organisms that contain them. Thus, the constraints on acceptable amino acids are **selective constraints** because unacceptable amino acid substitutions are eliminated by natural selection.

With synonymous substitutions, however, the selective constraints are weaker and, for some substitutions, may be negligible. One type of constraint on synonymous substitutions occurs through possible secondary structures that the RNA might form, in which certain nucleotides must undergo base pairing. A second, and perhaps more important, type of constraint occurs through codon preferences, which are correlated with the relative abundance of tRNA molecules that interact with and translate the codons. In bacteria and yeast, for example, highly abundant proteins tend to use codons for

abundant tRNA molecules, whereas proteins produced in small amounts tend toward codons for less abundant tRNA molecules (Ikemura 1985a,b). Substitution of a common codon by a rare codon for the same amino acid may not be neutral if it results in a lower rate of translation.

The rate of nucleotide substitution not only differs between nonsynonymous and synonymous sites, but it also differs in different regions of genes. Figure 17a illustrates the average rates of nucleotide substitution found in different parts of genes. The different parts are the 5′ (upstream) flanking region, the 5′ transcribed but untranslated region, the coding region (nonsynonymous substitutions only, with both the slowest and fastest rates shown), intervening sequences, the 3′ (downstream) transcribed but untranslated region, and the 3′ untranscribed region. On average, the fastest rates of substitution occur in intervening sequences and the 3′ flanking regions, but the average rates in the 5′ flanking regions and the 3′ untranslated region are all substantially faster than 0.88×10^{-9}, which is the average rate of nonsynonymous substitution in coding regions (see Figure 15a). Neutralists would argue that the high rates of substitution in noncoding regions, and variation among different parts of the coding regions, result from varying degrees of selective constraints on different parts of the gene. It is to be emphasized that Figure 17a depicts the averages of rates that are variable among genes; in individual instances, especially in comparisons of closely related species, there may be fewer substitutions observed in flanking sequences than in coding sequences, or fewer changes in synonymous sites than in nonsynonymous sites.

Comparison of nucleotide diversity in different functional regions of a single gene can reveal features of the gene's evolutionary history. For example, in 11 sequenced *Adh* genes of *Drosophila melanogaster* (Kreitman 1983), among 14 substitutions that were observed in the coding region, 13 were silent substitutions. Considering the genetic code (Table 1 in Chapter 3) and the codon usage in the *Adh* gene, it is possible to calculate the portion of the substitutions that would be silent if all substitutions occurred with equal frequency. This figure is about 30% in the case of the *Adh* gene in *Drosophila*, which implies that about 70% of the substitutions would be expected to cause amino acid replacements. Since only one out of 14 observed substitutions was an amino acid replacement, such substitutions are greatly underrepresented, which is consistent with the view that most amino acid replacements are eliminated from the population by purifying selection. This logic can be extended to argue that sequences that are conserved are likely to be functionally important; this type of reasoning led to the identification of a new open reading frame in the HIV (AIDS) virus genome (Miller 1988).

If an amino acid substitution does occur, its effect on the function of the protein product will depend on many factors, but one of the most important

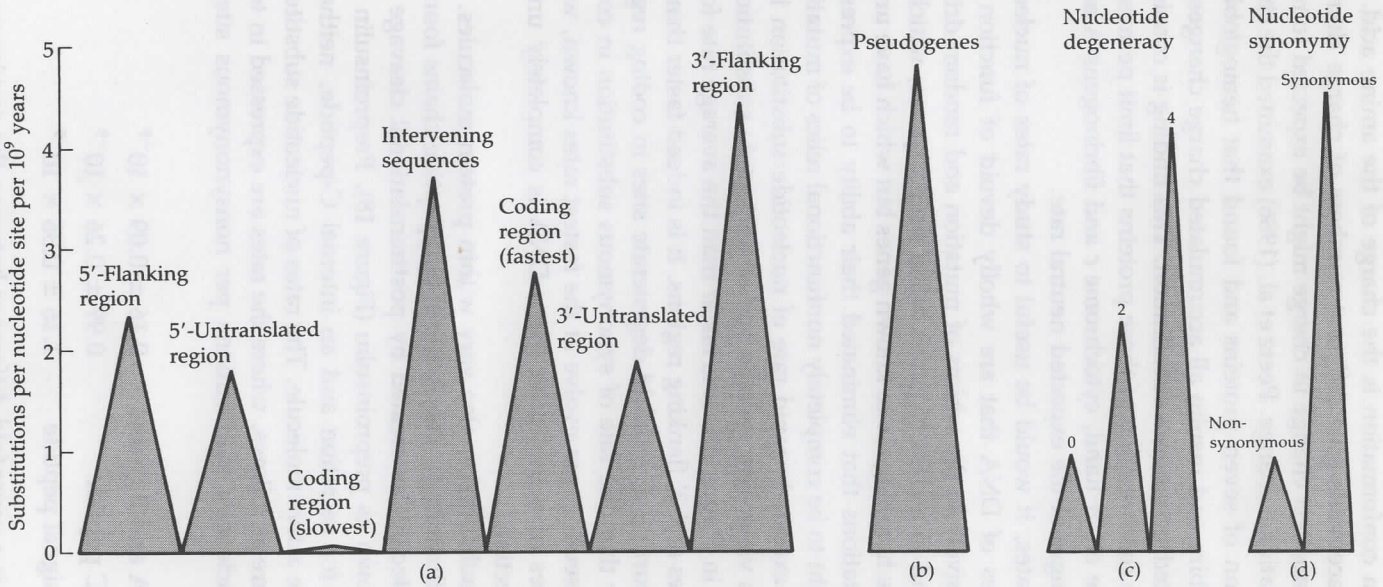


Figure 17. Comparison of average rates of nucleotide substitution in different parts of genes, and nucleotide sites. (a) Average substitution rates in different parts of genes. (b) Average substitution rate in pseudogenes. (c) Average substitution rates in nondegenerate (0), twofold degenerate (2), and fourfold degenerate (4) nucleotide sites. (d) Average rate of synonymous and nonsynonymous substitution. (Data from Li et al. 1985a.)

determinants of protein conformation is the charge of the amino acid. Different amino acid replacements give different numbers of charge changes, and in most cases the smallest change in charge might be expected to result in the smallest conformational change. Peetz et al. (1986) examined the charge changes in the evolution of seven proteins and found that hemoglobin α, hemoglobin β, myoglobin, and insulin all accumulated charge changes at a rate slower than expected by random substitution. This finding is consistent with constraints on the conformation of these proteins that limit permissible charge changes. On the other hand, cytochrome *c* and fibrinogens A and B accumulate charge changes at the expected neutral rate.

For comparison of rates, it would be useful to study rates of nucleotide substitution in stretches of DNA that are wholly devoid of function and therefore subject exclusively to the whims of mutation and random drift. A likely candidate is found in a class of genes called **pseudogenes**, which are DNA sequences that are homologous to known genes but which have undergone one or more mutations that eliminated their ability to be expressed. Pseudogenes are thought to be completely nonfunctional relics of mutational inactivation, and their extremely rapid rate of nucleotide substitution is offered in support of this view. The average rate of nucleotide substitution in pseudogenes is shown in Figure 17b. It is faster than the average rate found in intervening sequences and 3′ flanking regions. It is indeed faster than the rate of substitution occurring at fourfold degenerate sites in coding regions (Figure 17c), and faster than the rate of synonymous substitution in coding regions (Figure 17d). Pseudogenes evolve at the fastest rates known, which may correspond to rates of substitution when DNA is completely unconstrained by natural selection.

Rates of nucleotide substitution also vary within protein molecules. Human insulin is a good illustration. The A and B polypeptide chains found in the mature insulin molecule are created by posttranslational cleavage of a longer polypeptide known as preproinsulin (Figure 18). Preproinsulin contains a signal peptide for secretion and an internal C-peptide, neither of which are present in the active molecule. The rates of nucleotide substitution in these three regions are as follows, where the rates are expressed in terms of nonsynonymous nucleotide substitutions per nonsynonymous site per year (Li et al. 1985a):

Insulin A and B chains	$0.16 \pm 0.09 \times 10^{-9}$
Insulin C peptide	$0.99 \pm 0.26 \times 10^{-9}$
Insulin signal peptide	$1.16 \pm 0.36 \times 10^{-9}$

In insulin, while there is a sevenfold difference between the maximum and minimum rates of nonsynonymous substitution in different regions of the molecule, the rates of synonymous substitution differ only twofold. More-

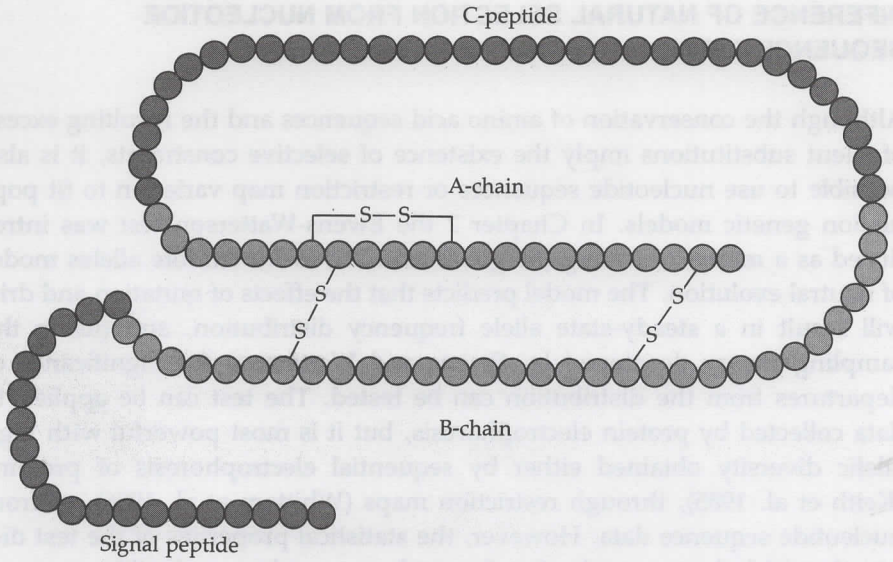


Figure 18. Functional specialization of regions within the insulin preproinsulin protein. The signal peptide functions in secretion, and the C peptide is cleaved from the molecule in maturation. The mature insulin polypeptide consists of the A and B peptides connected through disulfide bonds (—S—S—).

over, there is a negative correlation between functional importance and rate of nonsynonymous substitution within the insulin molecule. Many diverse amino acid sequences can serve as signal peptides provided they are hydrophobic, which suggests that selective constraints on signal peptides may be reduced in comparison with sequences in mature polypeptides. In insulin, as expected, the rate of nonsynonymous substitution is fastest in the signal peptide and slowest in the functional subunits of the mature molecule. This kind of negative correlation between selective constraint and substitution rate has also been observed in several other proteins (Li et al. 1985a).

From the selectionist viewpoint, while granting that substitutions in pseudogenes may be neutral and synonymous substitutions may be constrained by natural selection only weakly, it is nevertheless maintained that nucleotide substitutions that change amino acid sequences are inevitably subject to the action of natural selection of an intensity that is sufficient to counteract the effects of random genetic drift. Thus, selectionists would argue that amino acid substitutions that have occurred in a protein during the course of evolution became fixed by natural selection because they increased the fitness of the carriers through improvement in function of the molecule. However, neutralists argue back, the selectionist viewpoint cannot easily explain the negative correlation between functional importance and rate of substitution within proteins.

INFERENCE OF NATURAL SELECTION FROM NUCLEOTIDE SEQUENCE DATA

Although the conservation of amino acid sequences and the resulting excess of silent substitutions imply the existence of selective constraints, it is also possible to use nucleotide sequences or restriction map variation to fit population genetic models. In Chapter 2 the Ewens-Watterson test was introduced as a means of testing the goodness-of-fit to the infinite alleles model of neutral evolution. The model predicts that the effects of mutation and drift will result in a steady-state allele frequency distribution, and (using the sampling theory developed by Ewens and Watterson) the significance of departures from the distribution can be tested. The test can be applied to data collected by protein electrophoresis, but it is most powerful with high allelic diversity obtained either by sequential electrophoresis of proteins (Keith et al. 1985), through restriction maps (Whittam et al. 1986), or from nucleotide sequence data. However, the statistical properties of the test dictate that fairly large sample sizes be used, so complete nucleotide sequence data are not practical for this test at present. Using genes with many alleles, it may be possible to improve the power of the Ewens-Watterson test by analyzing the fit of the entire distribution of allele frequencies rather than the homozygosity (Clark 1988).

A test that is based on observed polymorphisms in the first, second, and third positions of codons has been explored by Sawyer et al. (1987b) using the nucleotide sequences of the *gnd* gene encoding 6-phosphogluconate dehydrogenase in a sample of seven natural isolates of *Escherichia coli*. In the sequenced region of the gene there were 12 amino acid replacements and 78 silent substitutions. All 12 of the amino acid replacements had a sample configuration of (6,1), meaning that 6 of the isolates had the same amino acid at the site, and only one had an alternative amino acid. Assuming a population model in which individuals die at a constant rate and are replaced by possibly mutant copies (the Moran model), the allele frequency distribution at steady state has a Dirichlet distribution in the case of equal mutation rates among all nucleotide types. The parameters of this distribution were fitted by maximum likelihood to the observed sample configurations, from which it was possible to show that the probability of observing 12 amino acid polymorphisms, all with (6,1) configuration, was less than 1%. The analysis also indicated that no more than half of the observed amino acid polymorphisms in *gnd* could be selectively neutral. Other models yielded the same qualitative conclusion that the *gnd* data do not fit the classical neutral model, but rather suggest that the amino acid polymorphisms are slightly detrimental.

Another test of neutrality makes use of nucleotide sequence data within

and among species (Hudson et al. 1987). The test is based on the idea that genes with a high rate of nucleotide substitution, as indicated by a large number of interspecific sequence differences, should also have a high level of intraspecific polymorphism. Sequences of at least two genes from a number of individuals of each of two species are needed to apply the test. Define S_i^A and S_i^B as the number of polymorphic nucleotide sites in gene i in species A and B, respectively, and D_i as the number of differences in gene i between a pair of alleles sampled randomly, one from species A and one randomly from species B. The expected values of these parameters are obtained from the infinite sites neutral model, assuming that the two species diverged t generations ago; that the population sizes are $2N$ and $2Nf$; and that each gene has an associated $\theta_i = 4N\mu_i$. Estimates of θ_i, f, and t are obtained by a least-squares method that gives the best fit of the expressions for the expected values and variances of S_i^A, S_i^B, and D_i to the data, and goodness-of-fit is tested with an appropriate χ^2 test. Using data from the *Adh* coding and 5′ flanking regions in *Drosophila melanogaster* and *D. sechellia*, Hudson et al. (1987) found that the observed values deviated significantly from the neutral model in a direction consistent with the operation of balancing selection acting on the coding region of *Adh*. This finding is consistent with Kreitman's (1983) observation of an excess of silent substitutions in *Adh*, except that the test of Hudson and co-workers makes use of the genetic variation observed within and among species.

For genes that determine surface antigens of pathogens or those that determine the major histocompatibility antigens of mammalian cells, the rates of nucleotide substitution can be quite high. One way to address whether the high rate of substitution is driven by selection is to examine the rates of synonymous and nonsynonymous substitution in these genes. For example, Hughes and Nei (1988) found that in the regions coding for the antigen recognition sites in the class I MHC (major histocompatibility) genes of humans and mice, the rate of nonsynonymous substitution exceeded the rate of synonymous substitutions by a ratio of 3:1. This ratio is the reverse of the usual situation (other regions in the same MHC genes show typical substitution rates, with an excess of silent substitutions). The excess of amino acid replacements is consistent with a model in which mutations that generate diversity are often advantageous, and hence natural selection accelerates the substitution process.

PHYLOGENETIC INFERENCE AND GENE GENEALOGIES

One of the most important results of the Moran model of genetic drift (Chapter 2) is that the stationary joint distribution of allele frequencies of two samples from a single population taken t generations apart is identical

to that of samples taken simultaneously from two populations of the same size after $(1/2)t$ generations of divergence. It is this reversibility of the Moran model that allows the construction of phylogenetic trees from measures of genetic divergence. If a gene in a pair of species or populations evolves in clocklike fashion, and if the degree of divergence between two genes implies that they have been diverging for t generations, then we can infer that the genes separated from a common ancestor $(1/2)t$ generations ago. This reasoning provides a group of methods of tree construction based on measures of genetic distance.

One such method is the unweighted pair-group method with arithmetic mean (UPGMA), or average distance method. Using a matrix of all pairwise distances, a tree is built up by first lumping the two species with the smallest distance. A new distance matrix is then constructed, with the lumped species now considered as one unit. If the lumped species were indexed i and j, then for all $k \neq i,j$ the distance from k to the cluster is $d_{k(ij)} = (1/2)(d_{ik} + d_{ij})$. In words, the distance from each species k to the cluster is the average of the distances from species k to each of the species in the cluster. The new distance matrix is again searched for the smallest element, and the appropriate lumping again occurs. This is repeated until all species are clustered into a tree. As simple as this method is, it has some very nice statistical properties, and computer simulations show that it yields the correct tree unless the evolutionary rates along branches differ greatly.

The average distance method constructs an **expected distance** tree, and assumes that the evolutionary time along each branch is proportional to the branch length. This is to be contrasted with a **realized distance** tree, which constructs branch lengths based on the actual number of substitutions, and is therefore better suited to situations where the substitution rate varies across the tree. One realized distance method is that of Fitch and Margoliash (1967). Suppose the number of substitutions distinguishing sequence i and j is d_{ij}. If the tree relating sequences 1, 2, and 3 has branch lengths A, B, and C (Figure 19), then the branch lengths can be estimated from

$$\begin{aligned} A &= (1/2)(d_{12} + d_{13} - d_{23}) \\ B &= (1/2)(d_{12} + d_{23} - d_{13}) \\ C &= (1/2)(d_{13} + d_{23} - d_{12}) \end{aligned} \tag{7.23}$$

These relations were found by solving the equations $d_{12} = A + B$, $d_{13} = A + C$, and $d_{23} = B + C$. With more than three sequences, the tree is built up by considering three units at a time, beginning with the two most closely related sequences and lumping the remaining sequences. If sequences 1 and 2 are the most similar, then the distances from sequence 1 to the remaining group is the average of the distances from sequence 1 to each member of the

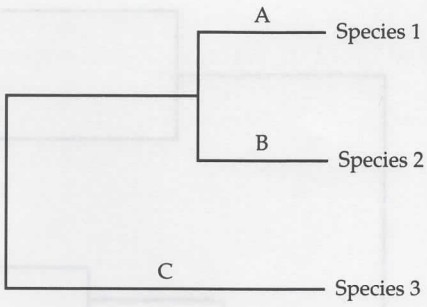


Figure 19. A simple phylogenetic tree. A, B, and C represent branch lengths from the most recent common ancestor.

group. In this way, only three distances are considered at a time, and Equations 7.23 allow branch lengths to be estimated.

There is no universal theory that provides a single optimal way to construct phylogenetic trees and, as basic as it seems, a distance matrix is not required by all methods. Another method, known as **maximum parsimony**, uses the smallest number of mutational events necessary to account for the evolution of a set of sequences from a common ancestor to construct the trees. There are a number of such parsimony methods based on trees with the smallest number of substitutions, but there is never a guarantee that the most parsimonious tree is the correct tree. For example, when rates of substitution differ in different branches of the tree, the parsimony method often fails to give the correct topology (Felsenstein 1978). Methods for constructing phylogenetic trees have been reviewed by Nei (1987), Fitch (1977), and Felsenstein (1981, 1982), among others. Massive simulation studies are also being done to test the statistical reliability of tree constructing methods (Rohlf and Wooten 1988; Kim and Burgman 1988; Sourdis and Nei 1988).

There is an important distinction between constructing trees using sequences from different species versus sequences from one species. The former yields a customary phylogenetic tree, while the latter produces what is called a **gene genealogy**. The relationships among species result from macroevolutionary processes, while allelic differences result from a number of microevolutionary processes including aspects of genetic transmission. Once the nucleotide sequences of alleles are known, the different alleles can be treated like genes in different species in applying standard methods for inferring a phylogenetic tree. However, great care is needed in constructing gene genealogies because recombination among the sequences results in a gross violation of the assumptions of most tree-building methods. Nevertheless, gene genealogies can be of great use in inferring the evolutionary history of a polymorphism. For example, they can reveal which of a group of alleles is older, or which alleles are more closely related to each other. Figure 20

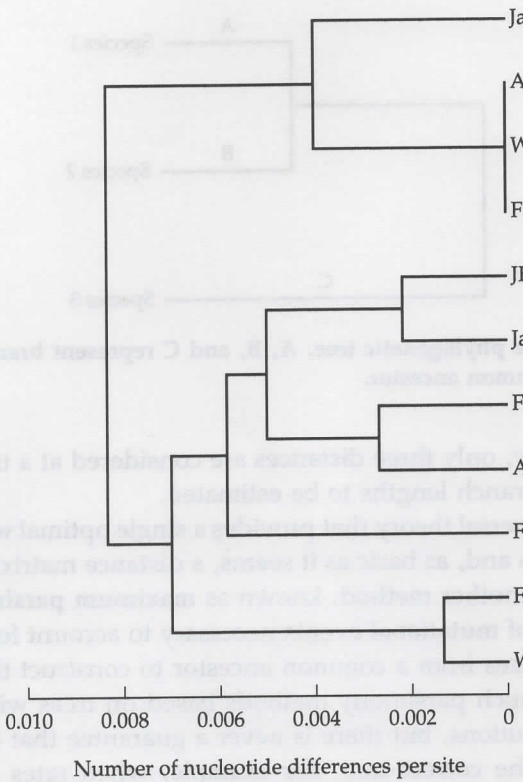
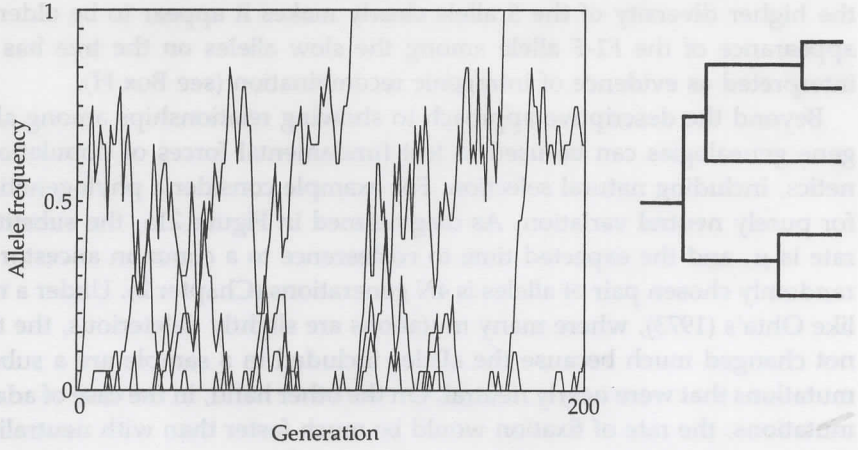


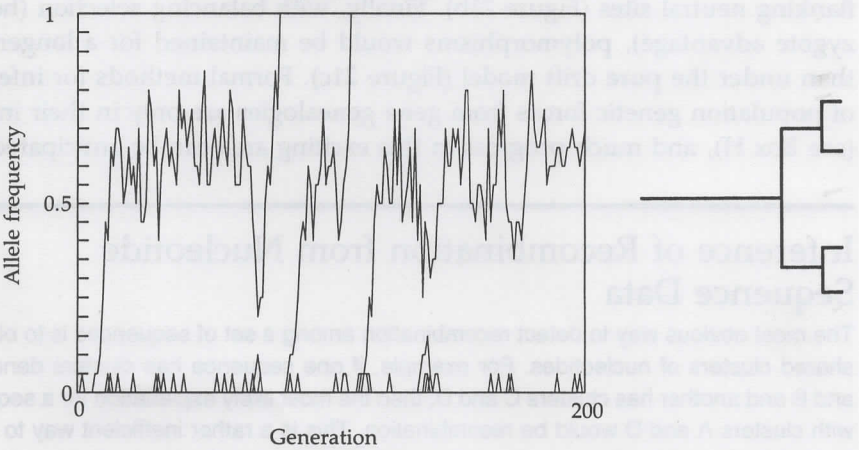
Figure 20. A UPGMA tree for 11 *Adh* alleles of *Drosophila melanogaster* based on 43 sequence differences. The scale is the number of nucleotide differences per site. Ja, Japan; Af, Africa; Wa, Seattle, Washington; Fl, Southern Florida; Fr, France. (From Stephens and Nei 1985.)

Figure 21. Computer simulations of the infinite alleles model of molecular evolution. In all three panels, $2N_e = 20$ and $\mu = 0.01$. (a) With strict neutrality, the expected time from mutation to fixation of alleles that will go to fixation is $4N_e$ generations and the time between successive fixations is $1/\mu$ generations. (b) Purifying selection (in this case with half of the mutations having a fitness of 0.5) results in less polymorphism at any given time. (c) Stabilizing selection (overdominance or frequency dependence) can retain alleles in a polymorphic state for much longer times. Representative trees are plotted to the right of each panel. ▷

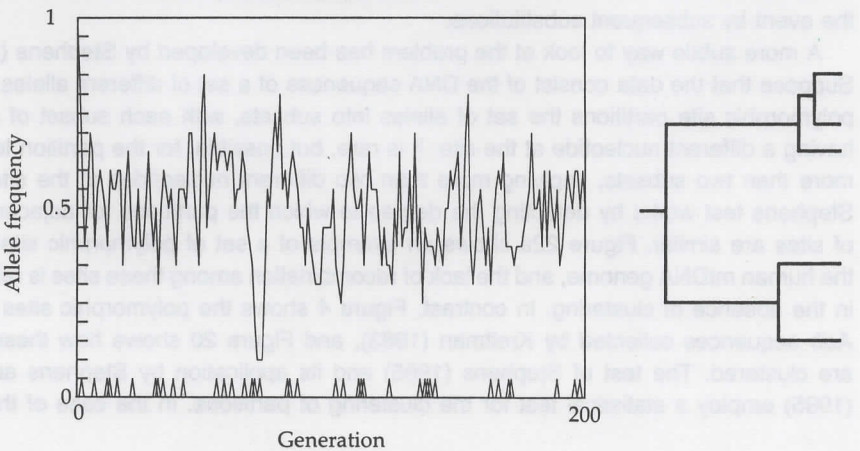
(a) No selection



(b) Purifying selection



(c) Stabilizing selection



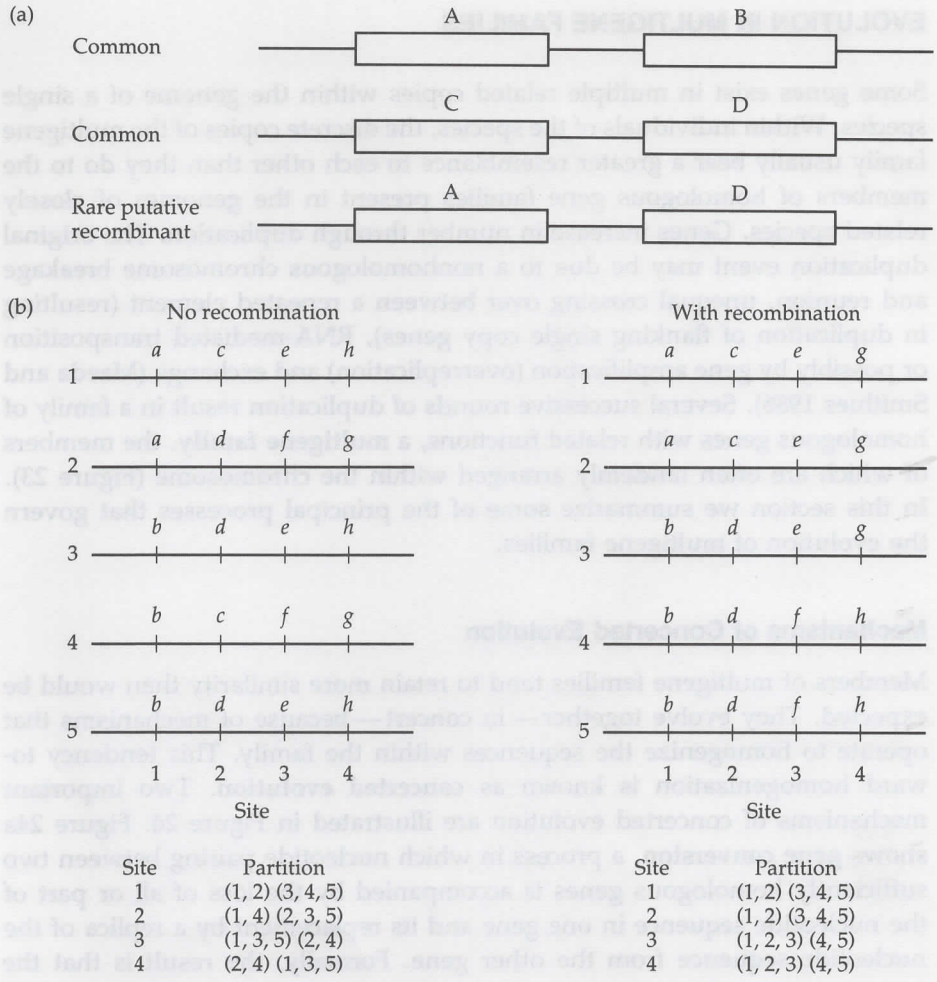
shows the gene genealogy from Kreitman's (1983) *Adh* sequence data, and the higher diversity of the *S* allele clearly makes it appear to be older. The appearance of the *F1-F* allele among the slow alleles on the tree has been interpreted as evidence of intragenic recombination (see Box H).

Beyond the descriptive approach to showing relationships among alleles, gene genealogies can be used to test fundamental forces of population genetics, including natural selection. For example consider a phylogenetic tree for purely neutral variation. As diagrammed in Figure 21a, the substitution rate is μ, and the expected time to coalescence to a common ancestor for a randomly chosen pair of alleles is $4N$ generations (Chapter 2). Under a model like Ohta's (1973), where many mutations are slightly deleterious, the tree is not changed much because the alleles included in a sample are a subset of mutations that were nearly neutral. On the other hand, in the case of adaptive mutations, the rate of fixation would be much faster than with neutrality, so that sites of adaptive mutation would have shorter coalescence times than flanking neutral sites (Figure 21b). Finally, with balancing selection (heterozygote advantage), polymorphisms would be maintained for a longer time than under the pure drift model (Figure 21c). Formal methods for inference of population genetic forces from gene genealogies are only in their infancy (see Box H), and much progress in this exciting area can be anticipated.

H Inference of Recombination from Nucleotide Sequence Data

The most obvious way to detect recombination among a set of sequences is to observe shared clusters of nucleotides. For example, if one sequence has clusters denoted A and B and another has clusters C and D, then the most likely explanation for a sequence with clusters A and D would be recombination. This is a rather inefficient way to detect recombination, since the sample must contain both parental sequences as well as the recombinant, and the recombination event must be relatively recent to avoid obscuring the event by subsequent substitutions.

A more subtle way to look at the problem has been developed by Stephens (1985). Suppose that the data consist of the DNA sequences of a set of different alleles. Each polymorphic site partitions the set of alleles into subsets, with each subset of alleles having a different nucleotide at the site. It is rare, but possible, for the partition to have more than two subsets, implying more than two different nucleotides at the site. The Stephens test works by detecting the degree to which the partitions for adjacent sets of sites are similar. Figure 22a shows an example of a set of polymorphic sites from the human mtDNA genome, and the lack of recombination among these sites is reflected in the absence of clustering. In contrast, Figure 4 shows the polymorphic sites of the *Adh* sequences collected by Kreitman (1983), and Figure 20 shows how these sites are clustered. The test of Stephens (1985) and its application by Stephens and Nei (1985) employ a statistical test for the clustering of partitions. In the case of the *Adh*



Site	Partition
1	(1, 2) (3, 4, 5)
2	(1, 4) (2, 3, 5)
3	(1, 3, 5) (2, 4)
4	(2, 4) (1, 3, 5)

Site	Partition
1	(1, 2) (3, 4, 5)
2	(1, 2) (3, 4, 5)
3	(1, 2, 3) (4, 5)
4	(1, 2, 3) (4, 5)

Figure 22. Inference of recombination from sequence data. (a) If haplotypes with blocks of sites A and B are common, and those with blocks of sites C and D are also common, then a haplotype with blocks A and D most likely arose by recombination. (b) Recombination affects the allelic partitions of sites. With no recombination, a null model fits the distribution of partitions, but recombination results in a departure from the model.

data, there is significant clustering, indicating that intragenic recombination (or gene conversion) has played a role in generating the observed sequence variation of this gene. Other methods exist to estimate the occurrence and number of recombination events that separate a set of sequences (Hudson and Kaplan 1985), and to estimate the importance of recombination in determining the joint distribution of allele frequencies (Hudson 1987).

EVOLUTION IN MULTIGENE FAMILIES

Some genes exist in multiple related copies within the genome of a single species. Within individuals of the species, the discrete copies of the multigene family usually bear a greater resemblance to each other than they do to the members of homologous gene families present in the genomes of closely related species. Genes increase in number through duplication. The original duplication event may be due to a nonhomologous chromosome breakage and reunion, unequal crossing over between a repeated element (resulting in duplication of flanking single copy genes), RNA-mediated transposition or possibly by gene amplification (overreplication) and exchange (Maeda and Smithies 1986). Several successive rounds of duplication result in a family of homologous genes with related functions, a **multigene family**, the members of which are often tandemly arranged within the chromosome (Figure 23). In this section we summarize some of the principal processes that govern the evolution of multigene families.

Mechanisms of Concerted Evolution

Members of multigene families tend to retain more similarity than would be expected. They evolve together — in concert — because of mechanisms that operate to homogenize the sequences within the family. This tendency toward homogenization is known as **concerted evolution**. Two important mechanisms of concerted evolution are illustrated in Figure 24. Figure 24a shows **gene conversion**, a process in which nucleotide pairing between two sufficiently homologous genes is accompanied by the loss of all or part of the nucleotide sequence in one gene and its replacement by a replica of the nucleotide sequence from the other gene. Formally, the result is that the

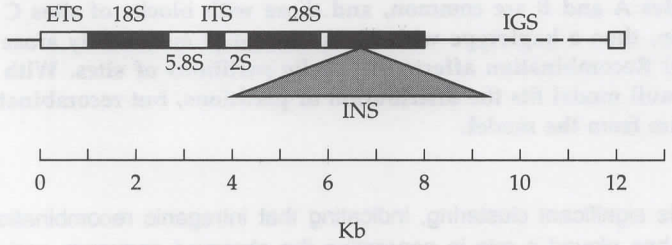
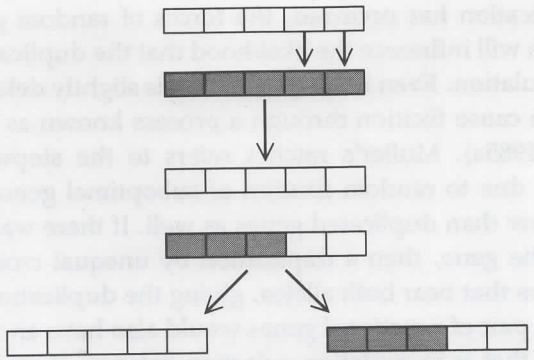


Figure 23. Organization of the rRNA genes in *Drosophila melanogaster*, a tandemly repeated multigene family. In addition to the 18S, 5.8S, 2S, and 28S rRNA genes, the repeated unit contains an external transcribed spacer (ETS), an internal transcribed spacer (ITS), and an intergenic sequence. Some copies also contain insertion sequences (INS).



(a) Gene conversion

(b) Unequal crossing over

Figure 24. Two mechanisms of concerted evolution that can homogenize the sequences of members of multigene families. (a) Gene conversion, in which genetic information in one gene is replicated and replaces genetic information in its homolog. (b) Unequal crossing over. The increase in the number of members of the gene family because of unequal crossing over can be corrected by a subsequent unequal crossing over, resulting in identical DNA sequences among several adjacent members of the multigene family.

sequence in one gene "converts" the sequence in the other gene to be exactly like itself.

Figure 24b illustrates concerted evolution by means of **unequal crossing over**. In tandem multigene families, misaligned pairing of the genes during meiosis can occur, and crossing over results in gametes that can have either an increase or a decrease in copy number. Unequal crossing over by itself does not result in a steady increase in copy number; rather it results in an increased *variance* in copy number. Thus, a subsequent crossing over can either correct the copy number or increase it further. Formally, the final result is genetically equivalent to gene conversion, because the nucleotide sequences of some members of the multigene family have been replaced with the nucleotide sequences of other members. At the same time that unequal crossing over increases the variance in copy number, it acts to homogenize the members of the family.

Once a duplication has occurred, the forces of random genetic drift and natural selection will influence the likelihood that the duplication will become fixed in the population. Even if the duplication is slightly deleterious, random genetic drift can cause fixation through a process known as **Muller's ratchet** (Charlesworth 1985a). Muller's ratchet refers to the stepwise decrease in maximal fitness due to random fixation of suboptimal genotypes; it applies to situations other than duplicated genes as well. If there was a heterozygote advantage for the gene, then a duplication by unequal crossing over could generate gametes that bear both alleles, giving the duplication an advantage. Duplication of a pair of functional genes would also have an initial advantage in a population that is in mutation–selection balance. If mutations occur on one or the other copy of the duplicated gene, and if it becomes necessary to have both copies to retain sufficient gene activity, then selective constraint continues to be important. If the duplication achieves fixation and there is no functional requirement to have two gene copies, then selective constraints have been relaxed. Depending on the balance of the forces of mutation, gene conversion, and unequal crossing over, the pair of genes may remain active and very similar, or they may diverge in function (such as different tissue-specific forms of amylase or LDH), or one gene may lose function and become a pseudogene.

Gene conversion and unequal crossing over operate to maintain similarity between copies of genes in multigene families. Theoretical studies of the processes have been carried out by Ohta (1982) and by Nagylaki (1984). In Ohta's model, a tandemly arranged multigene family consists of a fixed number n of members, and λ is the probability that a particular member of the gene family becomes converted by another member in any one generation. (Equivalently, λ is the probability of completion of a cycle of unequal crossing over resulting in the replacement of one gene in the family by another.) The mutation rate per gene is μ, and the effective population number is N. The probability that equal crossing over occurs within the gene cluster in any generation is given by β.

In this model there are three types of identity by descent that can occur between members of the multigene family, illustrated in Figure 25.

1. The probability c_1 that two genes in the same chromosome are identical by descent.
2. The probability c_2 that two genes in nonhomologous positions in different chromosomes are identical by descent.
3. The probability f that two genes in homologous positions in different chromosomes are identical by descent.

Complex formulas for the equilibrium values of c_1, c_2 and f have been derived by Ohta (1982), but they are greatly simplified when $4N\mu << 1$ and

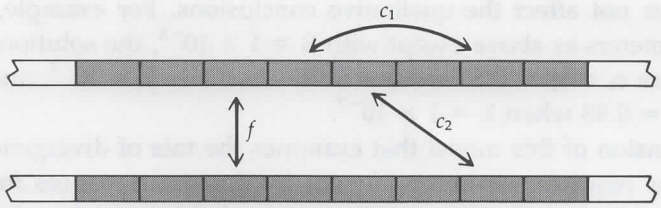


Figure 25. Three types of identity by descent in multigene families. They are the identity between genes at homologous sites (probability *f*), between genes at nonhomologous sites in the same chromosome (probability c_1), and between genes at nonhomologous sites in different chromosomes (probability c_2). (After Ohta 1982.)

when there is no equal crossing over within the gene cluster ($\beta = 0$). In this case, to a good approximation,

$$c_1 = c_2 = \frac{\lambda}{\lambda + (n - 1)\mu} \qquad \textbf{(7.24a)}$$

and

$$f = \frac{4N\lambda c_2 + 1}{4N\lambda + 4N\mu + 1} \qquad \textbf{(7.24b)}$$

Equation 7.24(b) is valid for all values of $4N\mu$. In this model, the quantity $n\mu$ equals the *total* rate of mutation in the entire multigene family, and this is very nearly $(n - 1)\mu$ when n is reasonably large. Thus, the implication of Equation 7.24(a) is that when the rate of gene conversion λ is substantially greater than the total mutation rate, the probability of identity by descent of nonhomologous genes within the family (c_1 and c_2) will be close to 1.0; and when λ is substantially smaller than the total mutation rate, the probability of identity by descent of nonhomologous genes will be close to zero. The random conversion process is analogous to that of random genetic drift in that if a single chromosome is followed through successive generations, the continual replacement of individual members of the linked family corresponds to the repeated sampling of gametes during reproduction.

The meaning of Equations 7.24(a) and 7.24(b) can be appreciated with a numerical example. Using the parameter values $n = 50$, $N = 5 \times 10^4$, $\lambda = 1 \times 10^{-5}$, and $\mu = 1 \times 10^{-8}$, the predicted values of c_1, c_2, and f can be calculated as $c_1 = c_2 = 0.95$ and $f = 0.968$. For $\lambda = 1 \times 10^{-7}$ the corresponding values are 0.17 and 0.98. With the larger rate of gene conversion, the identities differ little between genes at homologous (f) or nonhomologous positions (c_1, c_2). With the smaller rate of gene conversion, genes in homologous positions are much more likely to be identical than genes in nonhomologous positions. Incidentally, a small amount of equal crossing over in the gene

cluster does not affect the qualitative conclusions. For example, using the same parameters as above except with $\beta = 1 \times 10^{-3}$, the solutions to Ohta's formulas are $c_1 = c_2 = 0.88$ and $f = 0.92$ when $\lambda = 1 \times 10^{-5}$, and $c_1 = c_2 = 0.17$ and $f = 0.98$ when $\lambda = 1 \times 10^{-7}$.

An extension of this model that examines the rate of divergence between populations requires two more identity by descent measures for between-population identity. Tachida (1987) examined this model and found that unbiased gene conversion promotes differentiation at a single locus, but may either increase or decrease differentiation between loci, depending on population sizes and rates of mutation and conversion.

The joint effects of unequal crossing over and gene conversion in the presence of deleterious and advantageous mutations have been investigated by numerical simulation (Ohta 1987, 1988). The rate at which multigene family evolution proceeds was found to depend strongly on the rate of unequal recombination. Multigene families can avoid the accumulation of mutations when there is sufficiently strong natural selection, and positive selection is necessary for genes to evolve new functions. Walsh (1988) specifically addressed the question of genes within a family escaping from gene conversion, and hence being able to evolve new functions. Empirical examples of such escapes have been identified, such as the 18S rDNA pseudogenes in humans and apes (Brownell et al. 1983). Walsh supposed that the likelihood for conversion depended on sequence similarity, and his "k-hit" model supposes that conversion is completely stopped after a threshold of k mutations is present. With a per-nucleotide mutation rate of μ and a gene conversion rate between identical genes of λ, a gene can escape conversion if $2\mu/\lambda > 0.1$. This result was verified numerically, and it supports the intuitive notion that higher mutation rates and lower conversion rates lead to greater likelihood that a gene will escape conversion. The effects of biased gene conversion and selection on multigene families have also received attention (Slatkin 1986; Walsh 1986), and these models show that the effects of selection can be cumulative over the gene family, greatly augmenting the effect of selection.

Duplication and Divergence in Multigene Families

Duplicate genes can evolve in separate ways under the influence of natural selection, mutation, and random genetic drift. In time, some members of a multigene family may diverge to a greater or lesser degree in their function. This process of **duplication and divergence** is thought to be the major mechanism by which genes with novel functions are created.

Figure 26 illustrates the result of duplication and divergence in two related multigene families in humans: those coding for the β-like and α-like polypeptide chains of hemoglobin. The genes are specialized for different periods

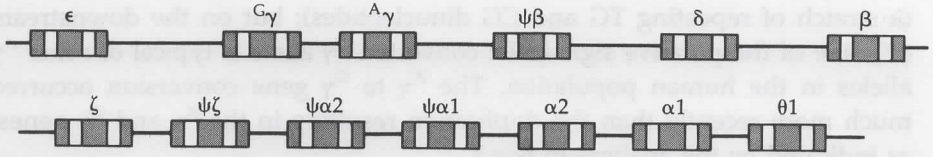


Figure 26. Organization of the β-globin and α-globin gene families in humans (located in chromosomes 11 and 16, respectively). The designation ψ denotes a pseudogene. The functional adult α-globin gene copies α1 and α2 are identical, and the fetal globin genes $^{G}\gamma$ and $^{A}\gamma$ are nearly identical. The genes ε and ζ code for embryonic hemoglobins. Judged by its nucleotide sequence, the α-like gene θ1 is potentially functional and may be expressed during restricted times or in restricted tissues during development. (Organization of the α-globin gene cluster after Marks et al. 1986.)

of life. The ε (epsilon) and ζ (zeta) genes are expressed in embryos; the $^{G}\gamma$ and $^{A}\gamma$ genes and the α genes in the fetus; and the β, δ, and α genes in the adult. The time of expression and function of the θ_1 gene is unclear (Marks et al. 1986).

The inference from differences in nucleotide sequence is that the original β–α duplication took place approximately 500 million years ago, when vertebrates were represented by the bony fishes, and the β–γ duplication took place about 80 million years ago, during the mammalian radiation. More recent duplications have also occurred, for example those leading to the two functional α genes, the cluster of three α-like pseudogenes, and the two γ genes. There are several models for the sequence of duplication, deletion, and conversion events that led to the current array of globin genes (Goodman et al. 1984; Hardies et al. 1984; Hardison 1984; Margot et al. 1988), but it appears well substantiated that the ancestral cluster that predated the mammalian radiation was 5′–ε–γ–η–δ–β–3′. With the mammalian radiation, the different orders of mammals evolved along different routes. In prosimian primates, such as lemurs, there was a fusion of η and δ. In higher primates, including humans, there was a δ–β conversion and γ duplication. In rodents, β and γ both duplicated, η was deleted, and there was a δ–β fusion, probably mediated by an unequal crossover. In rabbits, η was deleted and there was a δ–β conversion. Finally, in goats, γ was deleted, there was a δ–β conversion, and the remaining four-gene array was then triplicated!

The evolutionary history of the fetal globin genes in humans reveals that the $^{G}\gamma$ and $^{A}\gamma$ genes originated as part of a relatively recent 5-kb tandem duplication (Shen et al. 1981). Furthermore, evidence from nucleotide sequences strongly suggests that a gene conversion event also occurred that converted part of one particular $^{A}\gamma$ allele into a $^{G}\gamma$ allele (Slightom et al. 1980). The converted $^{A}\gamma$ allele is very similar to a $^{G}\gamma$ allele for about 1550 bp on the upstream (5′) side of a putative recognition signal for gene conversion

(a stretch of repeating TG and CG dinucleotides); but on the downstream (3′) side of the putative signal, the converted $^{A}\gamma$ allele is typical of other $^{A}\gamma$ alleles in the human population. The $^{A}\gamma$ to $^{G}\gamma$ gene conversion occurred much more recently than the duplication resulting in the $^{A}\gamma$ and $^{G}\gamma$ genes, as indicated by the analysis in Box I.

I Estimating the Time of a Duplication

The nucleotide sequence data from the $^{A}\gamma$ and $^{G}\gamma$ genes show that, on the upstream side of the putative gene conversion signal, 7 out of 1550 nucleotides differ between $^{G}\gamma$ and the converted $^{A}\gamma$ allele; but on the downstream side, 145 out of 1550 nucleotides differ. In both cases the majority of the 1550 nucleotides are noncoding. Use Equations 7.21 and 7.22 to estimate the amount of nucleotide substitution and its standard deviation on the upstream and downstream sides of the gene conversion signal. What evidence is there that the amounts are actually different in the two regions? Assuming that the rate of sequence divergence in the α-globin multigene family accumulates at the rate of 1% per 2.2 million years (an average based mainly on noncoding sequences), use the upstream data to estimate the time since the gene conversion event occurred, and the downstream data to estimate the time since the $^{G}\gamma$–$^{A}\gamma$ duplication occurred.

ANSWER With $\langle D \rangle = 7/1550$ differences upstream from the signal, $\langle k \rangle = 0.00453 \pm 0.0017$; and with 145/1550 differences downstream from the signal, $\langle k \rangle = 0.0999 \pm 0.0084$. The difference seems substantial, as its magnitude ($0.0999 - 0.00453 = 0.095$) is more than 11 times greater than the larger standard deviation. Assuming 1% sequence divergence per 2.2 million years, the time of the gene conversion is set at $0.00453 \times 100 \times 2.2 = 1.00$ million years; the time of the duplication itself is set at $0.0999 \times 100 \times 2.2 = 22.0$ million years.

The estimate of the time of occurrence of the $^{A}\gamma$–$^{G}\gamma$ duplication in Box I can be improved by using the nucleotide sequence data from the entire duplicated 5-kb region. In the entire region 14% of the nucleotide sites differ, which translates into $\langle k \rangle = 0.155 \pm 0.006$. This suggests a time for the duplication of $0.155 \times 100 \times 2.2 \times 10^{6} = 34$ million years (Shen et al. 1981).

In the regions of the $^{A}\gamma$ and $^{G}\gamma$ fetal globin genes that were unaffected by the gene conversion analyzed in Box I, there were 34 gaps, resulting from insertion and deletion events, among 4146 nucleotides (Shen 1981). The length distribution of the gaps, in number of nucleotides, was as follows:

Length	No.	Length	No.	Length	No.
1	12	6	1	13	1
2	3	7	2	23	1
3	2	8	0	122	1
4	4	9	0	Sum	34
5	6	10	1		

Assuming that each gap results from a single insertion or deletion event, we find that $L = 4146$ and $\langle D \rangle = 34/4146 = 0.0082$. These results imply that $\langle k \rangle = 0.00825 \pm 0.00142$. The rate of gap incorporation is therefore $0.00825/(2 \times 34 \times 10^6) = 0.121 \times 10^{-9}$ gap substitutions per nucleotide site per year. A rate of divergence in nucleotides of 1% per 2.2 million years equals $0.01/(2.2 \times 10^6) = 4.55 \times 10^{-9}$ nucleotide substitutions per nucleotide site per year. In the fetal globin genes, the rate of nucleotide substitution in noncoding regions is approximately 40-fold greater than the rate of incorporation of gaps.

Unequal crossing over in multigene families can result in a decrease in the number of genes as well as an increase (Figure 24). It is therefore not surprising that deletions of one or more of the hemoglobin genes are found in most parts of the world. Although usually very rare, in a few places the frequency of the deletions reaches levels too great to be accounted for by chance, especially in view of the observation that the carriers are mildly to severely anemic, or worse. For example, deletion of the β-gene results in death when homozygous, but it and other mutations that decrease the abundance of the β-hemoglobin chain are relatively common in the Mediterranean Sea basin, where malaria is endemic. For this reason, the decreased β-chain diseases are called β-thalassemias (literally translated as "sea anemias"). The well-established link between sickle-cell anemia and malaria (see Chapter 2) along with the geographical correlation between the β-thalassemias and malaria, provide a strong circumstantial case for malarial parasites being an important selective agent.

Deletion of one or more of the α-globin genes results in α-thalassemia. There are two α-genes (Figure 26), and deletion of one of them results in a mild form of anemia. For example, when a normal chromosome (α α) and a chromosome containing an α-deletion (- α) are combined in heterozygous condition (α α/- α), the individual is often indistinguishable from normal; on the other hand, the - α/- α homozygote exhibits a mild anemia.

The frequency of α-deletion chromosomes is significant throughout the malarial regions of the world, suggesting a causal link. Figure 27 strengthens the argument by demonstrating a direct correlation between the frequency of chromosomes containing the α-deletion and the prevalence of malaria. The study was carried out in Melanesia in the southwest Pacific, including the islands of Papua New Guinea, the Solomon Islands, Vanuatu, and New Caledonia (Flint et al. 1986). On the north coast of Papua New Guinea, where malaria is the most severely endemic, the frequency of the α-deletion chromosome reaches 68%. This figure may represent an approximate equilibrium in which the mildly harmful effects of the deletion are offset by the beneficial effects that result from increased resistance to malaria.

Red–green color blindness is a common X-linked disorder, with a frequency of about 8% in Caucasian males. The genes for the red and green

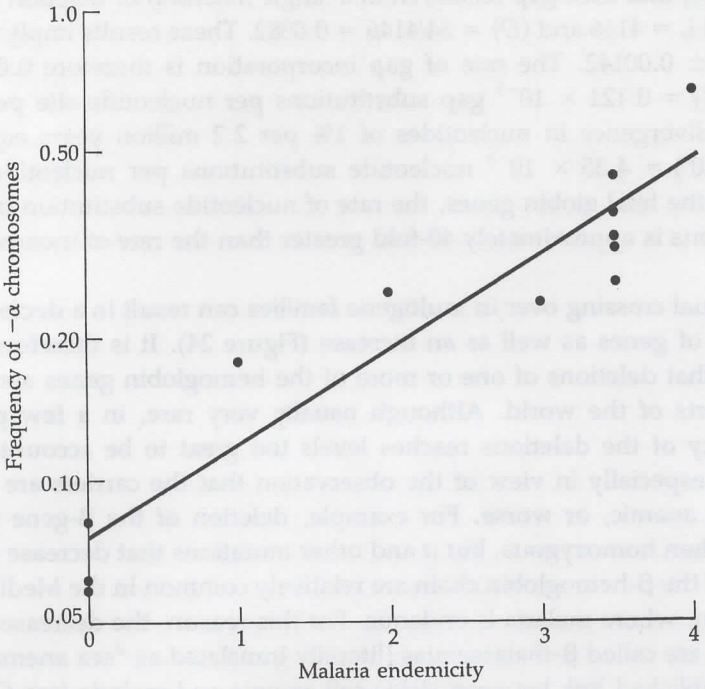


Figure 27. Association of α-thalassemia deletion (− α) with prevalence of malaria in Melanesia. The scale frequency is logarithmic, and that of malaria endemicity is proportional to the occurrence of the disease in different geographical regions. (Data from Flint et al. 1986.)

visual pigments have recently been cloned, and they match at 98% of their nucleotides, indicating that they arose by duplication. Individuals with normal color vision have one copy of the red pigment gene and varying numbers of copies of the green pigment gene. When genomic DNA from color blind males was analyzed by Southern blotting, those lacking green vision were lacking fragments of the green pigment gene. Further analysis showed that 24 of 25 color blind individuals had lost one or the other pigment gene through gene rearrangements that were either due to unequal crossing over or to gene conversion. In this example, the high sequence similarity of the red and green pigments works to our disadvantage by greatly increasing the likelihood of exchange events that lead to loss of color vision (Nathans et al. 1986).

Duplication of genes also occurs in plants, including a particularly important gene in plants that encodes the carbon fixing enzyme ribulose-1,5-bisphosphate carboxylase (RBC). The nuclear genome of many plants contains three genes encoding the small subunit of rbc, and Pichersky et al.

(1986) performed sequence alignments between the tomato and tobacco sequences for these genes. In the tomato, the *Rbcs-1* gene differs from the *Rbcs-2A* and *Rbcs-3A* genes at 13.9% and 13.1% of the nucleotide sites, respectively, while *Rbcs-2A* and *Rbcs-3A* differ at 10.7% of their sites. In comparison to the *RBC* gene sequence of tobacco, the genes *Rbcs-1*, *Rbcs-2A*, and *Rbcs-3A* of the tomato differ at 15%, 10.6%, and 11.3% of their nucleotide sites. Therefore, the *RBC* gene of tobacco appears to be more similar to the tomato *Rbcs-2A* gene than the pairwise comparisons among *Rbcs* genes within tomatoes. This finding is the opposite of concerted evolution, because there is an excess of within-species variation. Evidence from highly conserved regions of the RBC enzyme suggest a selective mechanism that accelerated divergence among genes within the tomato.

Ribosomal RNA Gene Family

Because rRNA constitutes the bulk of RNA in cells, it is relatively easy to isolate, and the gene family has been used extensively in inferring phylogenetic relationships (Woese 1981). The ribosomal gene cluster in *Drosophila melanogaster* consists of about 200 copies of a repeated unit on both the X and the Y chromosome, with each repeated unit containing an 18S and a 28S rRNA gene separated by an intergenic sequence (IGS) (Glover and Hogness 1977). The rRNA genes provide a clear example of concerted evolution because of great interspecific differences in spite of a high degree of sequence conservation within species (Coen et al. 1982). Furthermore, within individuals of *D. mercatorum* there appears to be little sequence variation, yet there are clear differences between individuals due to length variation in the intergenic sequence (Williams et al. 1985). This finding suggests the operation of a strong homogenizing force maintaining sequence fidelity within individuals. Among a series of 47 X and 47 Y chromosomes of *D. melanogaster* examined by Southern blotting of progeny from appropriate crosses, there appeared to be a polymorphism for a 5.1-kb insertion in the X-linked sequences only. The complete sequence of the intergenic sequence has been determined in *D. melanogaster*, *D. virilis*, and *D. hydeii*, and the sequences share a 220–240-bp repeat that is itself tandemly repeated. Figure 28 shows a block-plot comparison of the *D. melanogaster* intergenic sequences with themselves, and the process of slipped strand mispairing with the consequent generation of varying lengths of tandem arrays of repeats is immediately evident from the diagonal patterns.

An important aspect of rDNA in *Drosophila melanogaster* is that deficiencies in the gene cluster result in a phenotype recognized by short, thin bristles known as *bobbed*. Artificial selection on bristle number has in fact resulted in reduced rDNA copy number (Frankham et al. 1978). Bristle length and rDNA

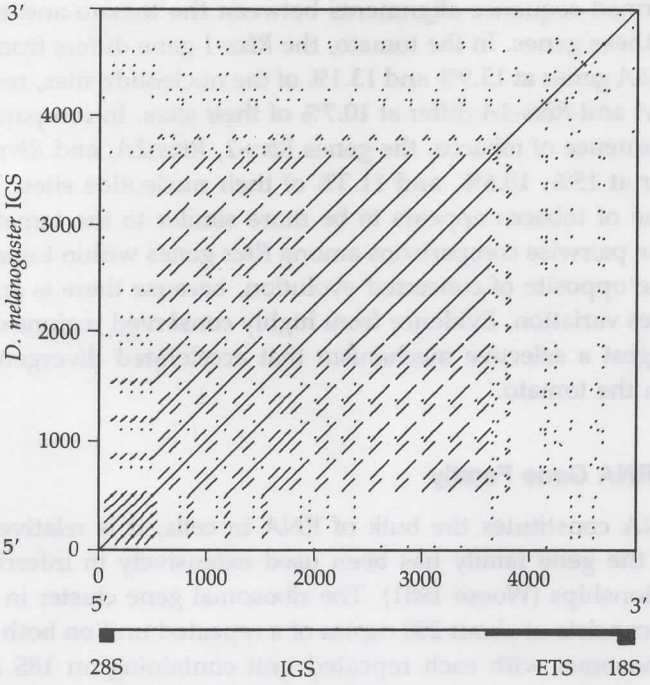


Figure 28. An identity block plot of the intergenic sequence of the rRNA gene cluster of *Drosophila melanogaster*. The sequence is compared to itself and sequences that are exact matches are indicated by diagonal lines. The repetition of the 220–240 bp sequence is evident. (Data from Tautz et al. 1987.)

content are associated, with a correlation coefficient of 0.79 (Dutton and Krider 1985), while complete loss of rDNA is lethal. *D. melanogaster* X chromosomes that have a reduced copy number (and cause the *bobbed* phenotype) gain copies when passed through males, by a process known as *magnification*. Komma and Endow (1987) showed that females bearing X–Y translocations also undergo magnification; Komma and Endow also demonstrated that a region of the long arm of the Y chromosome distinct from the rDNA repeats is necessary for magnification to appear. In males the rate of X–Y exchange is evidently low enough to retain differences between the X and the Y chromosome families (Williams et al. 1985, 1987), but the rate appears to be augmented when magnification occurs. Chromosomes that could only result from X–Y exchange events have been recovered from natural populations (Gillings et al. 1987). The rDNA mutants have been shown to be associated with retarded growth, and lines selected for fast and slow preadult development show consistent differences in the patterns of rDNA intergenic spacer

lengths (Cluster et al. 1987). Fast-developing lines tend to have a longer IGS with more copies of the 220-bp repeat, consistant with the observation that the 220-bp repeats have a transcription-enhancing function (Reeder 1984).

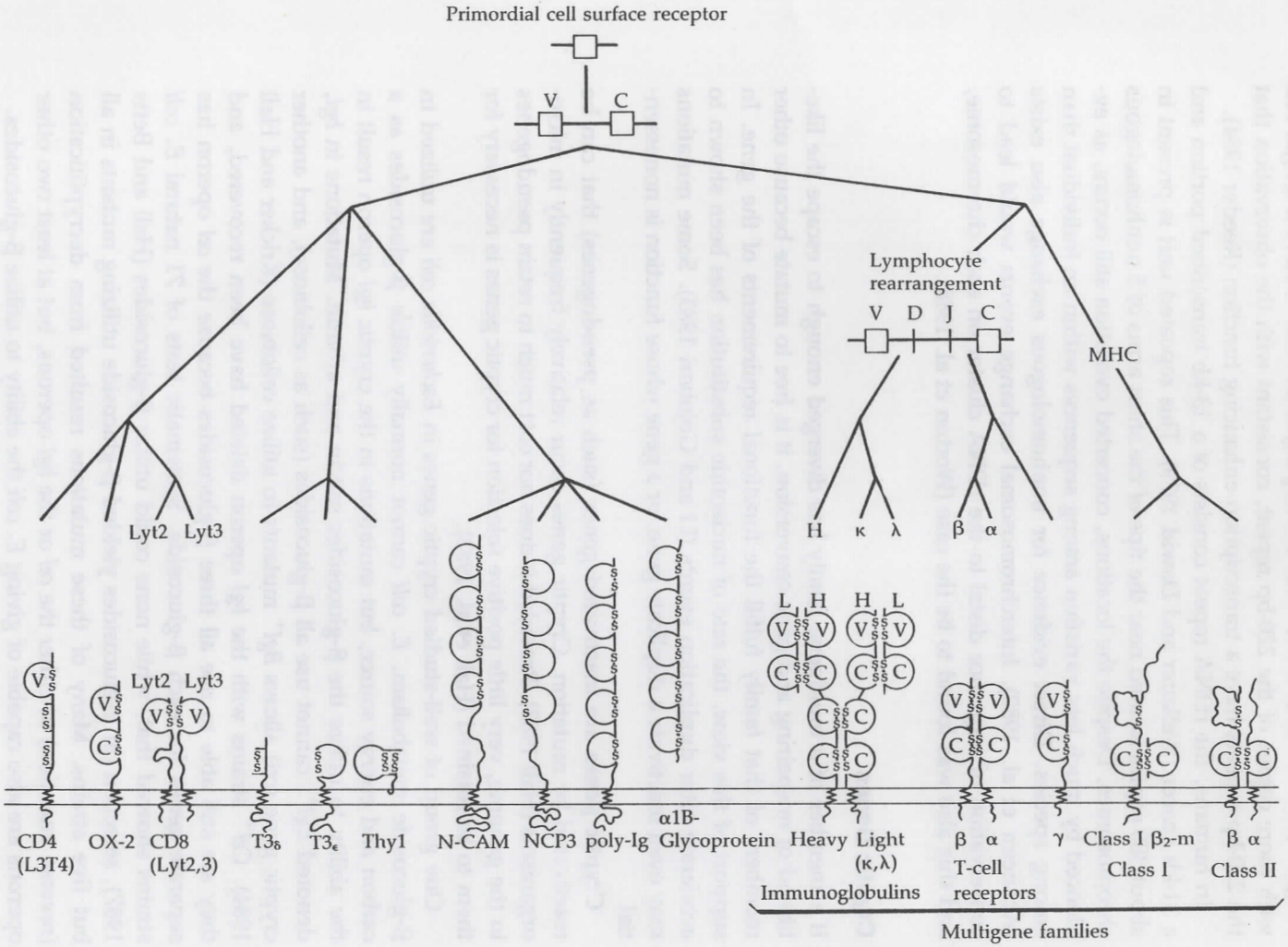
In humans, the rDNA repeat consists of a 13-kb transcribed portion and a 31-kb spacer (Wellauer and Dawid 1979). This repeated unit is present in about 300 copies located near the tips of the short arms of 5 nonhomologous chromosomes. Despite the locations, concerted evolution still occurs, as evidenced by much less variation among sequences within an individual than among species. Direct evidence for nonhomologous exchange also exists (Arnheim et al. 1980). Interchromosomal exchange events would lead to conservation of sequence distal to the rDNA cluster on each chromosome, and this also was found to be the case (Worton et al. 1988).

Cryptic Genes

If a member of a multigene family has diverged enough to escape the likelihood of mispairing and gene conversion, it is free to mutate because other members of that family fulfill the functional requirements of the gene. In support of this view, the rate of nucleotide substitution has been shown to accelerate after duplication events (Li and Gojobori 1983). Some mutations may even inactivate a duplicate gene, or a gene whose function is nonessential.

Cryptic genes are inactivated genes (such as pseudogenes) that can be reactivated by mutation. Cryptic genes occur relatively frequently in microorganisms (Hall 1983). Provided it does not cost much to retain pseudogenes in the genome, very little positive selection for cryptic genes is necessary for them to be retained (Hall et al. 1983).

One group of well-studied cryptic genes in *Escherichia coli* are utilized in β-glucoside metabolism. *E. coli* cannot normally utilize β-glucosides as a carbon and energy source, but mutations in the cryptic *bgl* operon result in the ability to utilize the β-glucosides salicin and arbutin. Mutations in *bgl*, denoted Bgl^+, cannot use all β-glucosides (such as cellobiose), and another cryptic gene (*cel*) allows Bgl^+ mutants to utilize cellobiose (Kricker and Hall 1984). Cel^+ strains with the *bgl* operon deleted have been recovered, and they are still able to use all three β-glucosides because the *cel* operon has separate genes for each β-glucoside. Systematic tests of 71 natural *E. coli* strains showed that, while none could utilize β-glucosides (Hall and Betts 1987), selection on β-glucosides yielded β-glucoside utilizing mutants in all but five strains. Many of these mutations resulted from decryptification (reverse mutation) of either the *cel* or the *bgl* operons, but at least two other operons are also capable of giving *E. coli* the ability to utilize β-glucosides.


Primordial cell surface receptor
V
C
Lymphocyte rearrangement
V D J C
MHC
Lyt2 Lyt3
H κ λ
β α
Lyt2 Lyt3
L H H L
CD4 (L3T4)
OX-2
CD8 (Lyt2,3)
T3δ
T3ε
Thy1
N-CAM
NCP3
poly-Ig
α1B-Glycoprotein
Heavy
Light (κ,λ)
Immunoglobulins
β α
T-cell receptors
γ
Class I β2-m
Class I
β α
Class II
Multigene families

Multigene Superfamilies

In some cases, several sets of multigene families and single-copy genes may share recognizable homology, implying a common ancestry, but they have undergone major divergence in function and relocation of position within the genome. These sets of historically related but functionally distinct genes constitute a **multigene superfamily**.

The remarkable similarities found among portions of genes in related gene families have suggested that many proteins have functional modules that can be combined in various ways in what is called **exon shuffling**. One example of this is tissue plasminogen activator (TPA), which has portions of three other proteins: plasminogen, epidermal growth factor, and fibronectin. What is striking is that the junctions of these protein segments fall precisely at intron–exon junctions. The epidermal growth factor shares exon similarity with several other proteins, including blood clotting factors IX and X, urokinase, and complement C9 (Doolittle 1985). The gene for the low-density lipoprotein (LDL) receptor in humans extends over 45 kb and contains 18 exons that show similarity to a bewildering variety of other proteins, including epidermal growth factor and blood clotting factors (Südhof et al. 1985). Just as a computer programmer recognizes the value of reusing subroutine modules in different programs, nature has capitalized on the efficiency of modular gene organization.

One extensively studied multigene superfamily that serves diverse functions in immunity is illustrated in Figure 29 (Hood et al. 1985; Hunkapiller and Hood 1986). The primordial single-copy gene may have coded for a cell surface receptor containing the basic homology unit of the superfamily — a unit that is about 110 amino acids in length, with a strategically placed

◁ **Figure 29. Proposed evolution of the immunoglobulin multigene superfamily from a primordial gene coding for a cell surface receptor. Details of the evolutionary relationships are speculative. The superfamily has diversified into 12 single-gene representatives (all of those at the left, plus β_2-microglobulin — β_2-m — at the right), and 8 multigene families (remaining representatives at the right). These include genes for antibodies, T-cell receptors, major histocompatibility antigens, and other functions. The single-gene members include T-cell molecules implicated in MHC recognition (CD4 and CD8) and possibly ion channel formation (T3δ, T3ϵ), an immunoglobulin transport protein (poly-Ig), a plasma protein (α_1B-glycoprotein), two molecules restricted to lymphocytes and neurons (Thy-1 and OX-2), two brain-specific proteins (N-CAM and NCP3), and β_2-microglobulin. The multigene families include the heavy (H) and light (κ, λ) components of antibody molecules, the α, β, and γ chains of T-cell receptors, and the Class I and Class II molecules from the major histocompatibility complex (MHC). (Adapted from Hood et al. 1985 and Hunkapiller and Hood 1986.)**

disulfide bridge and folding characteristics that enable it to combine with similar units. An early duplication and divergence of the primordial gene resulted in the variable (V) and constant (C) domains that have been so versatile in their diversification for specialized immune functions. In some members of the immunoglobulin superfamily, shown at the left in Figure 29, the functional products are usually individual polypeptide chains that sometimes contain internal duplications of the primordial folding unit. These products include the poly-Ig receptor that mediates the transport of immunoglobulin molecules across cell membranes.

In the other main branch of the superfamily, shown at the right, the functional products are usually aggregates of polypeptide chains. In this branch, there occurred multiple duplications of the V regions and specialization of D (diversity) and J (joining) regions during the evolution of the DNA splicing mechanism in lymphocytes, which today results in the tremendous diversity of antibodies and T-cell receptors. During the formation of heavy chain antibody genes in the lymphocytes, any one of a large number of DNA sequences coding for the variable part of the molecule can become spliced with any one of a small number of DNA sequences coding for the constant part, with diversity and joining regions incorporated in between. The plethora of possible V–D–J–C combinations enables enormous numbers of different possible antibodies to be formed, which is increased still further by slight variation in the exact positions of the splice junctions. An analogous type of splicing process occurs in the formation of antibody light chain genes and T-cell receptor genes.

Quite a different evolutionary pathway for formation of antibody heavy chain genes occurred in the primitive shark (Hinds and Litman 1986; Gilbert 1986). In the horned shark *Heterodontus francisci*, which diverged from mammals approximately 450 million years ago (Figure 11), the duplicated unit was the complete V–D–J–C unit. Although the sequences within each duplicated unit are joined by splicing, and some antibody variability results from variation in the splice junctions, the enormous number of possible V–D–J–C splicing combinations has been forfeited. As a result, the horned shark has a diminished repertoire of possible antibodies (Hinds and Litman 1986).

In yet another offshoot of the immunoglobulin superfamily, shown at far right in Figure 29, the C region underwent duplication and specialization to form molecules of the major histocompatibility complex (MHC), which, among other functions, are necessary for the T cells of the immune system to recognize foreign antigens.

Although many aspects of the immunoglobulin superfamily tree in Figure 29 are speculative, the molecules are undoubtedly related because comparison of the relevant units gives 15–40% homology at the amino acid level, and at the DNA level each homology unit is encoded in a separate exon. The

immunoglobulins thus demonstrate the immense evolutionary potential of repeated rounds of duplication and divergence through specialization of function.

Localized Highly Repetitive DNA Sequences

Most eukaryotic genomes contain tandemly arranged, highly repetitive DNA sequences near the centromeres or telomeres of chromosomes. In some cases the localized highly repetitive DNA accounts for a significant fraction of the genome. For example, in the American kangaroo rat *Dipodomys ordii*, more than half of the DNA consists of tandem arrays of variants of the simple sequences AAG, TTAGGG, and ACACAGCGGG, which comprise 2400, 2200, and 1200 million nucleotide pairs, respectively, and altogether account for more than 50% of the entire genome (Miklos 1985). The simple sequences are often not transcribed, and they are quite variable. For example, only about 25% of the sequences in the TTAGGG cluster in *Dipodomys* actually read TTAGGG; most of the others differ from TTAGGG in one or two nucleotides. Often called **satellite DNA**, localized highly repetitive DNA sequences provide abundant opportunities for gene conversion and unequal crossing over (Charlesworth et al. 1986).

The functions of localized highly repetitive DNA sequences, if any, are completely obscure. Miklos (1985) has classified functional hypotheses under four headings. Localized highly repetitive DNA sequences may be important for their role in the following:

1. Chromosome structure, for example, the determination of centromere strength;
2. Chromosome pairing and recombination;
3. Determining the three-dimensional architecture of the nucleus; and
4. Chromosome changes (genome reorganization), which may accompany the process of speciation.

An alternative hypothesis is that differences among organisms in localized highly repetitive sequences result in minimal phenotypic effects, and hence the sequences are not subject to the usual forces of natural selection acting at the level of phenotype.

Nonfunctional DNA sequences might have their own evolutionary dynamics determined by gene conversion and unequal crossing over, and the number of copies of each sequence could be amplified or diminished through unequal crossing over or other phenomena. When the evolution of DNA sequences is determined mainly by their own ability to replicate and undergo changes, and the course of events is affected minimally if at all by natural selection acting at the phenotypic level, the sequences are called **selfish DNA**

(Doolittle and Sapienza 1980; Orgel and Crick 1980). The molecular mechanisms that determine the evolution of selfish DNA (and of multigene families) are sometimes known collectively as **molecular drive** (Dover 1982).

All four functional hypotheses listed earlier may seem somewhat unattractive because of the variability that occurs in localized highly repeated DNA within species. For example, the genome of *Drosophila melanogaster* contains 20% satellite DNA, localized in the heterochromatic regions around the centromeres and composed of variants of the simple sequences AATAT, AATATAT, AAGAG, AAGAGAG, and AATAACATAG. However, approximately 3% sequence divergence in the satellites occurs among individuals, and up to 30% sequence divergence occurs in related species. Furthermore, the amount of heterochromatic DNA in *D. melanogaster* can be altered drastically without apparent effects on viability or fertility.

Functional hypotheses are also weakened by the sometimes peculiar location of satellite DNA in the genome. For example, approximately 60% of the genome in *Drosophila nasutoides* consists of satellite sequences, but—unlike *D. melanogaster*—all of the satellite DNA is localized to one giant chromosome (Figure 30). At the other end of the spectrum are species such as *D. littoralis* and *D. ezoana*, which lack detectable satellite DNA (Miklos 1985).

The chromosome reorganization hypothesis runs into difficulty because, while chromosome structure is sometimes drastically different between related species, this does not appear to be correlated with the amount of

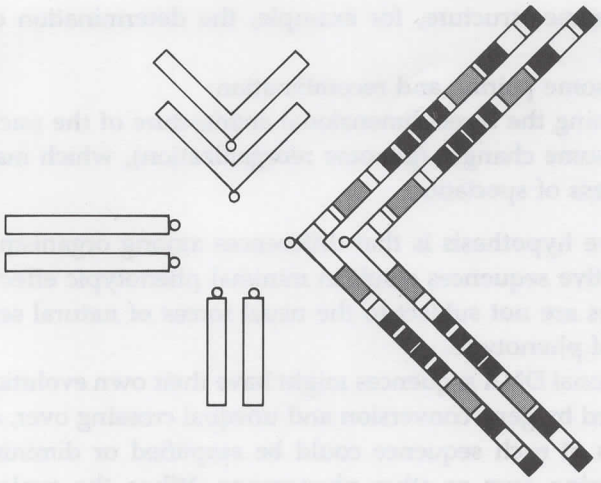


Figure 30. *Drosophila nasutoides* contains several localized highly repetitive DNA families (indicated by different shaded boxes), all of which are located in a single very large chromosome. (From Miklos 1985.)

Figure 31. Localized repetitive DNA sequence containing multiple GATA units (underlined) from the *Drosophila melanogaster* X chromosome. The arrows at the lower right define an ATG sequence in the opposite DNA strand which potentially initiates an open translational reading frame of 93 amino acids. The function of the sequence is unknown. (Data from Singh et al. 1984.)

satellite DNA. For example, chromosomes in the rat species *Rattus villosissimus* and *R. sordidus* have undergone extensive centromere rearrangement, with the result that many acrocentric chromosomes in the former species occur as metacentric chromosomes in the latter; yet the species contain negligible quantities of localized highly repetitive centromeric DNA. On the other hand, the *Drosophila* of Hawaii contain large amounts of satellite DNA, but the species are not differentiated by major chromosomal reorganizations.

The whole question of the function of localized highly repetitive DNA has been reviewed by Miklos (1985), who cites the inadequacies of all of the functional hypotheses. He concludes that satellite DNA is categorically selfish DNA, and that the principal mechanisms of its evolution are those of mutation, gene conversion, unequal crossing over, and random genetic drift.

One possible difficulty with the selfish DNA hypothesis is that, in some cases, simple-sequence DNA is transcribed in a developmentally specific manner. Figure 31 shows such a sequence from *D. melanogaster* located on the X chromosome. Note the many repetitions of the GATA motif (underlined). The DNA strand is transcribed into RNA from lower right to upper left, and the boldface region indicates a hypothetical product of 93 amino acids derived from translation of the RNA, starting with Met (codon AUG) at lower right (arrows). Related GATA sequences are found in the W sex chromosome in snakes and in the male-determining region of the Y chromosome in mice, raising the possibility that the sequence may play some unknown role in sex development (Singh et al. 1984).

Not all localized highly repeated DNA consist of variants of such short sequences as those in *Dipodomys ordii* and *Drosophila melanogaster*. The genome of the killer whale *Orcinus orca* contains a highly repeated sequence of 1579 bp, present in about 500,000 copies and accounting for approximately

15% of the total genomic DNA (Widegren et al. 1985). The sequence is organized in tandem repeats and is located in the heterochromatin of 8 of the 21 autosomes. Repeated sequences related to the one in the killer whale are found in all cetaceans (whales, dolphins, porpoises), testifying to the conservation of the sequence during evolution. Unfortunately, just as with highly repetitive DNA in other organisms, the function of the cetacean sequence is completely unknown. Indeed, it is not even certain that it has a function.

Dispersed Highly Repetitive DNA Sequences

A second major class of highly repetitive DNA in eukaryotes is not localized in clusters of tandemly repeating units, but is dispersed throughout the genome with single-copy sequences. In vertebrates, this dispersed highly repetitive DNA occurs in two categories, denoted SINEs and LINEs (Singer 1982).

SINEs (**S**hort **IN**terspersed **E**lements) are sequences typically shorter than 500 base pairs and occur in 10^5 or more copies in the genome. Like tRNA genes, they contain internal transcriptional start sites and are transcribed by RNA polymerase III. **LINEs** (**L**ong **IN**terspersed **E**lements) are sequences typically greater than 5000 bp and occur in 10^4 or more copies in the genome. They are processed pseudogenes (see page 404) and are transcribed by RNA polymerase II. Marked differences in the particular array of subfamilies of SINEs or LINEs or both are frequently observed among even closely related species. As in the case of localized highly repetitive DNA sequences, the mechanisms and possible significance of such massive and rapid changes in repetitive DNA in the genome are very obscure.

One example of a SINE in human DNA is the *Alu* family, so named because the sequence contains a characteristic restriction site for the restriction enzyme AluI. The *Alu* sequence is about 300 nucleotides in length, and members of the family show an average sequence divergence of about 15–20% from the *Alu* consensus sequence. *Alu* sequences are present in more than 500,000 copies in the human genome and constitute approximately 5% of the total DNA. Sequences closely related to *Alu* are found in other primates, and more distantly related sequences occur in rodents and probably in all placental mammals. The length of the rodent *Alu*-related sequence is only 130 nucleotides. The human *Alu* sequence consists of a head-to-tail dimer of the rodent sequence, with an additional 30 bp of DNA inserted into one of the monomers (Schmid and Shen 1985). Two randomly chosen human *Alu* sequences differ on the average at 15–20% of their nucleotide sites, which calculates to a time of divergence of between 16.7 and 23.3 million years. To the extent that gene conversion between *Alu* sequences is significant in

keeping the sequences more similar than they would otherwise be, the divergence times are underestimated. In the human genome there is an *Alu* element every 3–5 kb, but the distribution is not uniform. For example, the β-tubulin and thymidine kinase gene regions have about 10 times the average density of *Alu* repeats (Slagel et al. 1987), and *Alu* repeats show a preference for integrating into oligo-dA runs (Daniels and Deininger 1985).

The abundance of *Alu* sequences in the genome implies that many genes might have one or more *Alu* sequences in the same neighborhood. For example, the cluster of β-globin genes contains several copies of *Alu* sequences, which are of some significance in undergoing rare unequal crossing over resulting in hemoglobin disorders. A benign condition known as the hereditary persistence of fetal hemoglobin (high F disease) is associated with a deletion of the β and δ genes caused by unequal crossing over between two *Alu* repeats within the gene cluster; the reciprocal recombination product results in one form of β-thalassemia.

An example of LINEs in the human genome is the *Kpn* family of sequences (also called *LINE-1* or *L1*), where the designation *Kpn* refers to a characteristic restriction site. The *Kpn* sequences average about 2000 nucleotides, and the 50,000 copies of the sequence in the human genome account for about 4% of the total DNA. As with the *Alu* family, sequences related to *Kpn* are found in other mammals, including the mouse (Hardies et al. 1986) and the rabbit (Demers et al. 1986). Not all insertions of *L1* sequences are innocuous. Kazazian et al. (1988) found two cases of hemophilia A that were caused by *de novo* insertions of an *L1* sequence into exon 14 of the factor VIII gene whose function is necessary for normal blood coagulation. This insertional mutation event was evidently mediated by an RNA intermediate, and provides an obvious mechanism for natural selection to operate on *L1* elements. Another deleterious mutation caused by a transposable element in humans was an insertion of an *L1* sequence into the *myc* oncogene in a human breast cancer (Morse et al. 1988).

The abundance of dispersed repeated gene families is evident. Approximately 20% of the human genome consists of such sequences — 9% comes from the *Alu* and *Kpn* families alone. The profligacy of such sequences is matched only by our ignorance of their function. No function of highly repeated dispersed DNA sequences has yet been demonstrated. As with the localized highly repeated sequences, the possibility must be considered that the dispersed class of sequences is essentially functionless selfish DNA.

Some clues to the origin and evolution of LINE sequences come from analysis of their molecular structure. In molecular organization they resemble processed RNA molecules. At the 5′ end they contain transcriptional start signals, at the 3′ end they often contain many consecutive adenine nucleotides (poly-A tail), and they contain no intervening sequences. Dispersed

repeated DNA sequences also feature a short, direct (not inverted) duplication of 10–20 nucleotides at the termini.

In their molecular organization, LINE sequences strongly resemble a class of pseudogenes known as **processed pseudogenes**. Processed pseudogenes are thought to result from the reverse transcription of an RNA molecule into DNA, followed by insertion of the DNA into the genome. The reverse transcription and integration process can be carried out by an enzyme called **reverse transcriptase**, which is coded in the genome of a class of RNA-containing viruses called **retroviruses**. In cells infected with retrovirus, the reverse transcriptase makes a DNA copy of the viral RNA, and another enzyme inserts the DNA into the chromosome. When reverse transcription and integration happen to a processed RNA molecule, the result is a dispersed duplicate copy that is generally transcriptionally inactive due to loss of regulatory sequences. Such a sequence is known as a processed pseudogene. Many genes are known to have processed pseudogene counterparts, including the genes for human λ-immunoglobulin and β-tubulin, rat α-tubulin and cytochrome *c*, and mouse α-globin. Not all genes that have been processed through an RNA intermediate are pseudogenes. The human phosphoglycerate kinase gene (*PGK*) occurs as an active X-linked gene, a processed X-linked pseudogene, and an autosomal gene with remarkable properties. The normal *PGK-1* gene contains 11 exons and 10 introns, but the autosomal gene has no introns and has remnants of a poly-A tail, strongly implying that it was reverse transcribed from an RNA transcript. The intron-free autosomal gene (*PGK-2*) is expressed in human testes (McCarrey and Thomas 1987).

One hypothesis for the origin of LINE sequences is that they originate from genes that are actively transcribed at times in development that are particularly favorable for reverse transcription and integration (Schmid and Shen 1985). This hypothesis has the virtue that the family of sequences could be maintained even if only a small minority of the members were capable of transcription on their own. And indeed, most dispersed repeated sequences are not transcribed, or are transcribed only because of accidental read-through transcription from upstream promoters. In this hypothesis, as pseudogene copies of the sequence are lost through deletion and other processes, new ones are created by reverse transcription and integration of transcripts from functional members of the gene family. Walsh (1985) has estimated that, at equilibrium, 1.7–15% of the members of a repeated gene family may retain their ability to function.

The processed pseudogene model of dispersed repeated DNA evolution is illustrated in Figure 32 (Hardies et al. 1986). The functional, transcribed copies of the gene family are shown at the top, and the dotted lines represent gene conversion that promotes concerted evolution of the functional genes.

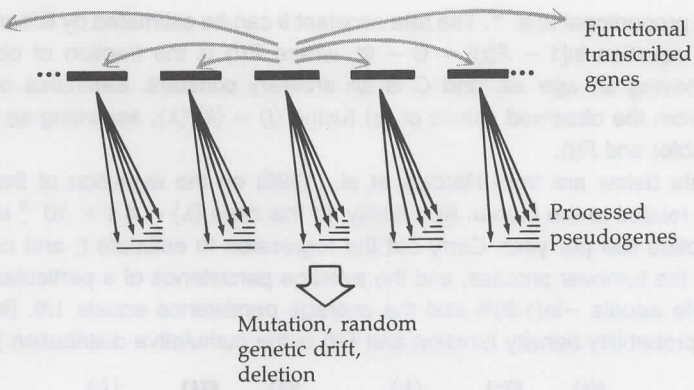


Figure 32. Model for the evolution of a dispersed highly repetitive family of processed pseudogenes. A small number of functional genes (top), which undergo concerted evolution by means of gene conversion, are transcribed under conditions that favor reverse transcription and integration into numerous dispersed chromosomal locations. The resulting nonfunctional genes undergo mutation and random genetic drift, and are ultimately eliminated by deletion or other mechanisms. (From Hardies et al. 1986.)

The gene in the center is a preferred donor for gene conversion (**biased gene conversion**). Emanating from the functional genes are numerous copies of processed pseudogenes distributed throughout the genome. These copies are essentially functionless and undergo sequence divergence promoted by mutation and random genetic drift, which is offset in part by gene conversion and other homogenizing processes among the pseudogenes. Eventually the pseudogene sequences are cleared from the genome by deletion, extreme sequence rearrangement, or divergence.

One implication of the model in Figure 32 is that eventually a balance is reached at which the clearance of old pseudogenes from the genome is equaled by the creation and insertion of new ones. In the equilibrium state there is a steady turnover among sequences in the family, but the total number neither grows nor shrinks. Studies of a dispersed repeated sequence in the mouse related to human *Kpn* suggest a turnover with a half-life of approximately two million years. That is, after two million years, half the members of the gene family will have been removed and replaced with new ones. However, the *Kpn* family may evolve more rapidly than is typical.

Rate of Disappearance of Repetitive DNA Sequences

J

If the disappearance of dispersed repetitive DNA sequences from the genome is like radioactive decay, in that a constant fraction θ of the sequences in the genome are eliminated per unit time, then the fraction of sequences present at time 0 that still exist

at time t is proportional to $e^{-\theta t}$. The rate constant θ can be estimated by linear regression using the equation $\ln[1 - F(t)] = C - \theta t$, where $F(t)$ is the fraction of observed sequences having an age $\leq t$, and C is an arbitrary constant. Estimates of θ can be obtained from the observed values of $\langle k \rangle$ (using $\langle t \rangle = \langle k \rangle / \langle \lambda \rangle$, assuming an estimate of λ is available) and $F(t)$.

The data below are from Hardies et al. (1986) on the evolution of the rodent *L1* sequence related to the human *Kpn* family. In this case $\langle \lambda \rangle = 4.1 \times 10^{-9}$ substitutions per nucleotide site per year. Carry out the regression to estimate t, and calculate the half-life of the turnover process, and the average persistence of a particular sequence. The half-life equals $-\ln(1/2)/\theta$ and the average persistence equals $1/\theta$. (In the table, $f(t)$ is the probability density function and $F(t)$ is the cumulative distribution.)

$f(t)$	$F(t)$	$\langle k \rangle$	$f(t)$	$F(t)$	$\langle k \rangle$
0.312	0.312	0.00334	0.061	0.819	0.02027
0.196	0.508	0.00670	0	0.819	0.02370
0.031	0.539	0.01007	0.027	0.846	0.02715
0.126	0.665	0.01345	0.061	0.907	0.03062
0.093	0.758	0.01685	0.061	0.968	0.03410

Values of $\langle t \rangle$ corresponding to $\langle k \rangle$ are, in the same order, 0.814×10^6, 1.63×10^6, 2.46×10^6, 3.28×10^6, 4.11×10^6, 4.94×10^6, 5.78×10^6, 6.62×10^6, 7.47×10^6, and 8.32×10^6. The slope of the regression of $\ln[1 - F(t)]$ on t estimates θ; in this case $\langle \theta \rangle = 3.40 \times 10^{-7}$ per year. The half-life is 2.04×10^6 years and the average persistence of a sequence is 2.94×10^6 years.

Hardies et al. (1986) have emphasized that the model of sequence evolution in Figure 32 allows ample scope for the occurrence of natural selection at the level of phenotype, not only through the direct effects of the functional members of the family, but through indirect effects of the pseudogene copies. The very abundance of the pseudogenes implies that many unrelated genes may have pseudogenes in the same vicinity, as is the case with *Alu* sequences interspersed in the β-globin cluster. Some fraction of these linked pseudogenes may alter the level, timing, or tissue distribution of transcription of the genes to which they are linked, or they may have subtle effects on chromatin structure that affect gene expression. Through any of a diversity of mechanisms, pseudogene copies of dispersed highly repeated gene families could, in principle, have effects on phenotype and thus be subject to the influence of natural selection. While true in principle, such effects have not yet been demonstrated. To the extent that such effects can safely be ignored, the evolutionary mechanism of highly dispersed repeated DNA sequences is that of selfish DNA, subject to the conflicting forces of neutral mutations and random drift and the diverse homogenizing processes of concerted evolution.

TRANSPOSABLE ELEMENTS

DNA sequences that can change their location within the genome are called **transposable elements**. The process of transposition requires a protein, called transposase, which is usually coded within the sequence of the transposable element itself. Many transposable elements undergo transposition through a replicative process with DNA or RNA intermediates. In most cases, when transposition to a new location occurs, one copy of the transposable element remains behind in its original location. Transposable elements can thereby increase in copy number in the genome. Most transposable elements are also able to regulate their own rate of transposition. Generally speaking, the rate of transposition is regulated in such a way that the rate of transposition per element decreases as the number of copies of the element increases, which prevents the occurrence of runaway transposition. Regulation of transposition is thought to be an evolved characteristic rather than being intrinsic to the biochemistry of the process. Several different types of transposable elements can be distinguished by their nucleotide sequence organization or by the details of their mechanisms of transposition or regulation.

Factors Controlling the Population Dynamics of Transposable Elements

Transposable elements were originally discovered in maize through genetically unstable mutations resulting from insertion near the affected genes. They are now known to be ubiquitous among prokaryotes and eukaryotes. Insertions of transposable elements can in some cases alter the regulation of genes in their vicinity, which prompted their original designation in maize as **controlling elements** (Fedoroff 1984; McClintock and Keller 1983). The propensity of transposable elements to transpose and increase in copy number reveals a dynamic, fluid aspect of the genome not previously recognized. Furthermore, many transposable elements are related to viruses, or they use other mechanisms that enable their dissemination among organisms or even among species. Transposable elements are thus examples of selfish DNA. Realistic models for the population dynamics of transposable elements usually incorporate several features.

1. A rate of infection, in which genomes previously lacking the transposable element become infected with it.
2. A rate of transposition, which determines how rapidly the copy number increases; the effects of regulation are taken into account by assuming that the rate of transposition is a decreasing function of copy number.
3. A mechanism, or combination of mechanisms, for eliminating elements

from the population, otherwise the copy number would increase indefinitely. The usual assumption is that the presence of transposable elements in the genome decreases fitness, resulting in the elimination of some elements by means of natural selection, or that elements can be eliminated from the genome by means of genetic deletion.

Through the study of such models, the diversity and novel attributes of transposable elements are being incorporated into the concepts of population genetics (e.g., Langley et al. 1983; Montgomery and Langley 1983; Charlesworth and Charlesworth 1983; Kaplan and Brookfield 1983; Sawyer et al. 1987a; Sawyer and Hartl 1986; Hartl and Sawyer 1988).

The model of Langley et al. (1983) makes the following assumptions. First, transposable elements are selectively neutral. Second, the rate of transpositions varies as $1/n$, where n is the number of elements in the genome. Third, when transposition occurs, the element inserts into a site that is not found in any other individual in the population (analogous to the infinite sites model). Fourth, elements are deleted at a rate of μ per generation. Fifth, sampling takes place at the level of individuals, with binomial sampling of $2N_e$ individuals. Sixth, there is no linkage disequilibrium among inserts (due to sufficiently high recombination). Finally, it is assumed that the transposable element is never completely lost by stochastic sampling, which implies that elements enter the population by a low rate of immigration. This model leads to the same frequency spectrum as the infinite alleles model for one locus — the expected number of sites bearing a transposable element insertion having a frequency between x and $x + \delta x$ is

$$\Lambda 4N\mu x^{-1}(1 - x)^{(4N\mu-1)}\delta x \qquad (7.25)$$

where Λ is the expected number of transposable elements per haploid genome at steady state. Brookfield (1986) took the theory a step further in predicting properties of sequence divergence within transposable elements.

One simple prediction is that if transposable elements have a recessive deleterious effect at the site of insertion, the density of transposable elements on the X chromosome would be decreased owing to the hemizygous state in males. Although a significant deficit of the element *412* was found on the X chromosome of *D. melanogaster*, the elements *roo* and *297* were not deficient in X-linked copies (Montgomery et al. 1987). Other models suggest the importance of meiotic recombination in copy number regulation, and predict the accumulation of transposable elements in genomic regions having depressed recombination (Charlesworth 1985b). The data that will allow further assessment of the theories is just becoming available.

One potentially important consequence of transposable elements is the resulting increase in mutation. Whether the increased mutation can result in

increased genetic variance that is of significance to quantitative traits was tested by Mackay (1984, 1985). She selected for increased or decreased numbers of abdominal bristles in populations of *Drosophila melanogaster* that were either dysgenic or nondysgenic, and found a greater response in the dysgenic lines. (The dysgenic lines had rates of transposition of the transposable element *P*, as described later in the chapter.) Although the results have proven difficult to repeat, they do provide an initial suggestion that *P* element insertion may yield significant increases in heritability of this quantitative trait. Another potentially important consequence of the presence of transposable elements is the opportunity for unequal crossing over that they provide. For example, Davis et al. (1987) showed that duplication and deficiency chromosomes in *D. melanogaster* could be explained by unequal exchanges among a set of three *roo* element inserts near the *white* gene.

Insertion Sequences and Composite Transposons in Bacteria

Bacteria contain several types of transposable elements, the simplest of which are **insertion sequences**. Insertion sequences are transposable sequences, typically about 1000–2000 nucleotides in length, that contain short nucleotide sequences repeated at each end in inverted orientation, and at least one long open translational reading frame thought to code for the transposase protein. A typical type of insertion sequence from intestinal bacteria, denoted IS*50*, is diagrammed in Figure 33a. IS*50* is 1500 nucleotides in length and produces

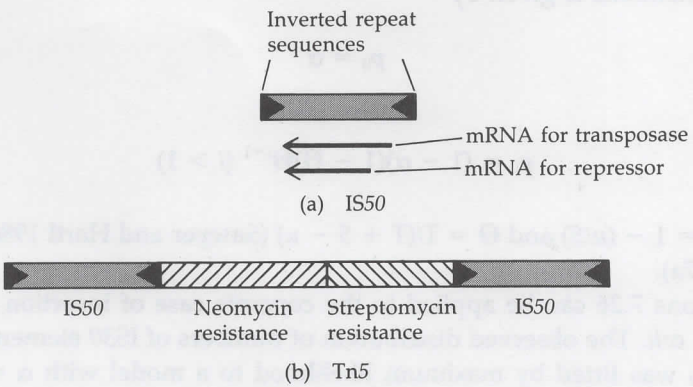


Figure 33. Examples of two types of transposable elements in bacteria. (a) Insertion sequence IS*50*, which is flanked by short inverted repeats and codes for its own transposase protein and repressor. (b) Composite transposon Tn*5*, which contains two copies of IS*50* flanking a central region containing genes for neomycin and streptomycin resistance.

two overlapping mRNA molecules, translated in the same reading frame, one for the transposase protein and the other for a repressor protein which regulates the rate of transposition. The intestinal bacterium *Escherichia coli* contains more than six well-characterized insertion sequences, the numbers of which have been determined among 71 natural isolates (Sawyer et al. 1987a). These distributions have been used to estimate the parameters in population models.

Models of transposable elements in *E. coli* are greatly simplified because of asexual reproduction, a low rate of recombination among strains, and a low rate of deletion of insertion sequences. The "state" of a strain may be described in terms of the number of copies n of an element it contains, and in accord with the controlling factors discussed in the previous section, let

1. u equal the rate at which uninfected cells become infected (that is, go from $n = 0$ to $n = 1$);
2. T equal the rate of transposition in infected strains [that is, the rate at which cells in state n ($n > 1$) go to state $n + 1$]; and
3. S equal the decrease in fitness of infected cells relative to uninfected cells.

The most general models of this type allow T and S to be functions of n, but here we will assume that they are constant. However, assumption (2) does incorporate regulation of transposition because the rate of transposition per element in a strain equals T/n, which is a decreasing function of n.

With these assumptions, the population attains an equilibrium distribution of numbers of elements, in which the probability p_i that a strain contains exactly i elements is given by

$$p_0 = \alpha \tag{7.26a}$$

and

$$p_i = (1 - \alpha)(1 - \Theta)\Theta^{i-1} \ (i > 1) \tag{7.26b}$$

where $\alpha = 1 - (u/S)$ and $\Theta = T/(T + S - u)$ (Sawyer and Hartl 1986; Sawyer et al. 1987a).

Equations 7.26 can be applied to the concrete case of insertion sequence IS*30* in *E. coli*. The observed distribution of numbers of IS*30* elements among 71 strains was fitted by maximum likelihood to a model with $\alpha = 1/2$ and $\Theta = 1/2$. With these parameters the distribution simplifies to the remarkably simple formula $p_i = (1/2)^{i+1}$ for $i > 0$. Among 71 strains, therefore, the expected numbers of strains containing i elements are as follows (the observed numbers are also given):

No. elements	Expected no. strains	Observed no. strains
0	35.5	36
1	17.8	16
2	8.9	13
3	4.4	2
4	2.2	2
≥5	2.2	2

The strains with five or more elements have been grouped in order to carry out a χ^2 test of goodness-of-fit. This χ^2 test has three degrees of freedom because α and Θ were estimated from the data. The value of χ^2 equals 3.48, which has an associated probability level of about 0.35. Thus, the simple model for IS*30* fits the observed data very well. Although the χ^2 test cannot be completely trusted in this case because of the small expected numbers in some of the categories, the conclusion is supported by a more exact statistical test (Sawyer et al. 1987a). In an attempt to explain the observed distributions of *IS* elements in *E. coli*, Sawyer et al. (1987a) fitted the distributions to a variety of models having different degrees of transposition regulation. The distribution of IS*1* fits Equations 7.26 with $\alpha = 1/5$ and $\Theta = 5/6$; IS*2* fits the same model with $\alpha = 2/5$ and $\Theta = 2/3$; and IS*4* fits the same model with $\alpha = 2/3$ and $\Theta = 3/4$. In all cases, the χ^2 values indicate an excellent fit to the model. It is generally not difficult to obtain an adequate fit of a two-parameter distribution to an observed distribution, so the observation of a good fit does not prove that this is the correct model. However, the observation of strong differences among elements emphasizes that the various elements face different levels of copy number regulation.

It should not be assumed that insertion sequences and other transposable elements are necessarily always harmful to their hosts. Box K demonstrates that the insertion sequence IS*50* can result in an advantage in fitness in *E. coli* cells cultured in chemostats.

Examples Where an Insertion Sequence and a *Tn* Element Appear to be Selectively Favored

K

Strains of *Escherichia coli* bearing the transposable element Tn*10* have an elevated rate of mutation, and in competition experiments, Chao et al. (1983) found that Tn*10*-bearing strains sometimes outcompeted strains lacking Tn*10*. This effect was found to be frequency-dependent, in that Tn*10* did not win in competition if its initial frequency was too low. Chao et al. argued that the low initial frequency resulted in a lower chance for beneficial mutations to occur in the Tn*10*-bearing strains. All of the Tn*10*-bearing strains that increased in frequency in chemostats were found to have an insertion of

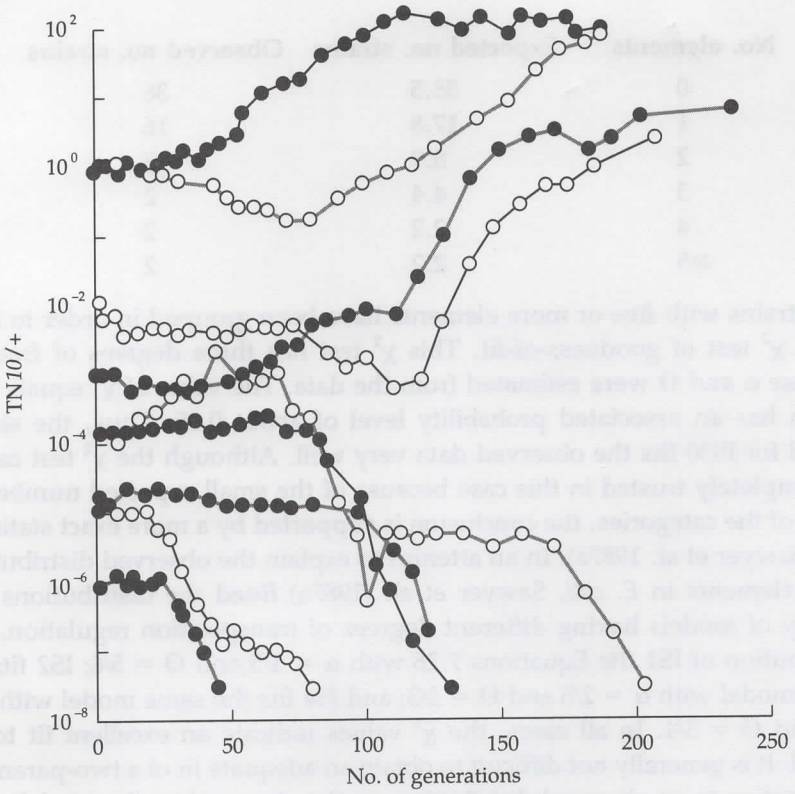


Figure 34. When chemostats are started with an initial frequency of Tn*10*-bearing strains greater than about 10^{-4}, the Tn*10* strains outcompete those that lack the transposon. This selective advantage is not seen if the initial frequency is less than 10^{-4}. It is thought that a certain minimal frequency of Tn*10* is needed to be assured that favorable mutations occur in the Tn*10* bearing strain. Although the Tn*10* strains have a higher mutation rate, favorable mutations will more likely occur in the non-Tn*10* strain if Tn*10* is rare enough, simply because there are so many more of the non-Tn*10* bearing cells. (From Chao et al. 1983.)

IS*10* near a particular location in the genome. Elimination of this IS*10* insertion resulted in a return to normal fitness in these strains, suggesting that an IS*10* insertion in this particular site is beneficial (Figure 34). The physiological reason why an IS*10* insertion at this site increases fitness is unknown.

The same outcome with a different mechanism was obtained for insertion sequences IS*50* by Hartl et al. (1983). Two strains of *E. coli* were subjected to competition for glucose in chemostats. The strains were genetically identical except that strain A contained one copy of the insertion sequence IS*50*. The selection coefficient favoring the IS*50*-bearing strain was calculated from the following data:

Hours of competition	Percent strain A
0	56.2
14	71.5
24	81.0
32	86.3
50	94.5

The selection coefficient per hour favoring strain A is estimated as the slope of the regression of ln(A/B) against time (in hours), where A and B are the relative frequencies of strains A and B in the chemostat. The logarithms, from top to bottom, are 0.249, 0.920, 1.450, 1.840, and 2.844; the slope of the regression is 0.052 per hour. This is a very large selective advantage, but it persists for only about 50–75 hours in the chemostat. The mechanism of the selective advantage is unknown, but, unlike the Tn*10* case, it does not result from transposition (Hartl et al. 1983).

Aside from their own evolutionary dynamics, insertion sequences are important because they can mobilize other sequences in the genome. When two copies of an insertion sequence are on flanking sides of an unrelated sequence, the inverted repeats that are used in transposition may be those at the extreme ends. This kind of insertion sequence "sandwich" constitutes a composite transposable element, or **transposon**, which transposes as a single unit. Transposon Tn*5*, an example of a transposon mobilized by IS*50*, is illustrated in Figure 33b. The central sequence of Tn*5* contains two regions coding for resistance to the antibiotics neomycin and streptomycin. Possession of Tn*5* would obviously confer a major selective advantage on a bacterial cell in an environment containing either or both antibiotics. In the case of Tn*10* (Box K) the flanking IS element is called IS*10*.

Mobilization of genes for antibiotic resistance, heavy-metal resistance, and other functions is one of the principal evolutionary implications of transposable elements in bacteria. They enable the piecewise assembly of specialized, infectious plasmids. **Plasmids** are autonomously replicating circular molecules of DNA that exist in bacterial cells, which in many cases contain genes that promote their transfer between different organisms. Many plasmids contain genes, such as genes for antibiotic resistance, that are highly advantageous to their hosts in certain environments. These genes are often contained in transposons, and they undoubtedly entered the plasmid through transposition from a different plasmid or from the genome of a previous host. Infectious plasmids containing multiple antibiotic-resistance genes are called **resistance transfer factors**. Transposon Tn*5* (Figure 33b) was originally found in a resistance transfer factor in a species related to *Escherichia coli*. Testifying to the ability of transposons to get around is the fact that the streptomycin-resistance gene in Tn*5* is not expressed in intestinal bacteria but in the soil bacterium *Rhizobium*.

Transposable Elements in Eukaryotes

Transposable elements can have important genetic consequences as mutagenic agents, through effects on the expression of genes they become inserted into or near. Their consequences in population genetics and evolution are less well understood. Transposable elements of several different types are widespread in both animals and plants. For example, *Drosophila melanogaster*

contains multiple copies of each of 50–100 different transposable elements (Spradling and Rubin 1981; Rubin 1983). Although few of these elements have been studied in detail from the standpoint of population genetics, indirect evidence suggests that most of the elements, like insertion sequences in bacteria, are mildly harmful to the host (Golding et al. 1986).

Two of the most thoroughly studied types of transposable elements are diagrammed in Figure 35. Although both examples occur in *D. melanogaster*, other transposable elements with analogous types of sequence organization occur in other organisms. Figure 35a illustrates the *P* element, in which the coding region in the middle is flanked by short inverted repeated sequences. The coding region of many *P*-like elements is interrupted by one or more intervening sequences, but some *P*-like elements contain a single uninterrupted reading frame (for example, the *Drosophila* element *mariner* discussed in Hartl 1988). In the case of the *P* element, transposition of the element is normally restricted to cells in the germ line because, in somatic cells, the third intron in the transcript of the *P* element is not excised (Laski et al. 1986). This mechanism was conclusively demonstrated by constructing a *P* element in vitro that lacks the third intron. Since somatic tissues can perform the other steps of transcript processing, when this engineered *P* element is introduced into fly embryos by microinjection, it now transposes in somatic tissue at a very high rate (Laski et al. 1986). *P* elements have also been found in members of the *Drosophila willistoni* group, including *D. nebulosa*. The sequence similarity of the elements from *D. melanogaster* and *D. nebulosa* is 94% (Lansman et al. 1987), consistent with the model that *D. melanogaster* may have recently acquired *P* elements from *willistoni* species in areas where their ranges overlap. However, the *D. nebulosa P* element is not able to provide transposase function for *D. melanogaster P* elements, nor does the *D. melanogaster P* element mobilize the *D. nebulosa* element. The degree of se-

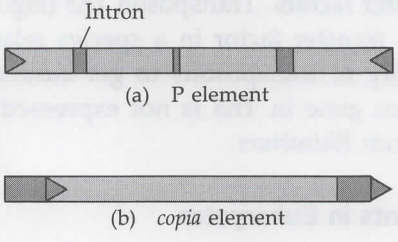


Figure 35. Examples of two types of transposable elements in *Drosophila*. (a) The *P* element, flanked by short inverted repeats and containing four exons interrupted by three introns. (b) The *copia* element, a retrovirus-like element, which contains a single uninterrupted reading frame between long direct repeats.

quence conservation of both *P* elements and *copia*-like elements across species is consistent with interspecific horizontal transmission between reproductively isolated species (Stacey et al. 1986).

Figure 35b illustrates the sequence organization of the *copia* transposable element. *Copia*-like elements are flanked by long direct repeats (that is, not inverted), and their coding region often contains a single uninterrupted open reading frame. *Copia* exemplifies transposable elements that are closely related to, or identical with, the RNA retroviruses discussed earlier, which reverse-transcribe their RNA into DNA and integrate it into the genome. Much remains to be learned about the regulation of transposition of *copia*. In a study by Biémont et al. (1987), after 69 generations of sibmating, 16 lines showed no change in *copia* positions, while one line showed a complete reshuffling of genomic locations of *copia*. The mechanism for these transpositional bursts is unknown.

The next section focuses on a phenomenon occurring in *D. melanogaster* that results from the presence of the *P* element but serves more generally to illustrate the potential importance of transposable elements in evolution. The evolutionary implications of retroviral-like elements are discussed in the following section.

Hybrid Dysgenesis in *Drosophila*

Severe genetic abnormalities that occur in the germ line of certain hybrids of *Drosophila melanogaster* are referred to collectively as **hybrid dysgenesis** (Kidwell et al. 1977b; Engels 1983, 1988). The abnormalities include partial or complete sterility in either or both sexes, increased mutation, chromosome breakage, and nondisjunction. These diverse phenomena result from mobilization of the *P* element in the germ line of the hybrids (Bingham et al. 1982).

The occurrence of hybrid dysgenesis is related to the distribution of the *P* element among strains. The element occurs in multiple copies in most strains of *D. melanogaster* recently collected from natural populations, but it is absent from old laboratory strains. Recently collected strains containing *P* elements are called P strains, and old laboratory strains and others lacking the *P* element are called M strains. Hybrid dysgenesis occurs when females of an M strain are crossed with males of a P strain.

Hybrid dysgenesis does not occur in the reciprocal cross of P females with M males because of substances that P females transmit through the egg cytoplasm (Engels 1979). The nature of the suppressing cytoplasm, known as P cytotype, and its relation to the genomic *P* elements themselves remains obscure (Engels 1983). However, recent evidence suggests that the *P* element codes for its own repressor of transposition, and that maternal transmission

of the repressor accounts for P cytotype (Engels 1988). The specificity of transposition of *P* elements by *P–M* hybrid dysgenesis is seen in the failure of *P–M* dysgenesis to induce transposition of any of 19 other families of transposable elements in *D. melanogaster* (Eggleston et al. 1988).

Although the hybrid animals in hybrid dysgenesis exhibit reduced fertility, the mobilization of *P* element in the germ line means that most of the chromosomes transmitted by the hybrids will have become infected with *P* elements. Thus, provided that the reduction in fertility in the hybrids is not too great, the *P* element can spread throughout the entire population. This conclusion has been confirmed for *P* elements both theoretically (Uyenoyama 1985) and experimentally (Kiyasu and Kidwell 1984; Scavarda and Hartl 1987). Apparently, the *P* element is a relative newcomer to the genome of *Drosophila melanogaster*, having invaded the genome just in the past 50 years since the old laboratory strains were collected (Anxolabéhère et al. 1988). Although most samples from natural populations are of the P cytotype, there is a very unusual cline in P and M cytotype frequencies on the east coast of Australia (Boussy and Kidwell 1987). Although the simplest theory predicts a spreading of the P cytotype, the cline has apparently not moved during the period from 1983 to 1986. Since another example of hybrid dysgenesis in *D. melanogaster* is well documented (Bregliano and Kidwell 1983; Bucheton et al. 1984), which results from the mobilization of a transposable element completely unrelated to the *P* element, it appears that the infection, spread, and turnover of transposable elements may be a very rapid and frequent process, at least in this species.

Endogenous Retroviruses in Mammals

Mammals and some other organisms contain, as a normal constituent of their cellular DNA, multiple copies of the genome of several unrelated classes of retroviruses. DNA sequences that are homologous to retroviruses and that have become fixed in the genome of a species are known as **endogenous retroviral sequences**, or **virogenes**. Several distinct types of endogenous sequences occur in the vertebrate genome. Present in 10–100 copies per genome, each type of virogene had its origin in reverse transcription and integration of a different RNA retrovirus. Under suitable conditions of cell culture, endogenous viruses can give rise to mature retrovirus particles. Although not ordinarily infectious in cells of their species of origin, the released particles are infectious in cells of related species. The released viruses are typical retroviruses in that a small number of reverse transcripts of the viral RNA become incorporated into the host cell DNA. These observations and others suggest that endogenous retrovirus sequences are advantageous to the host in conferring some degree of immunity to viral infection (Benveniste 1985).

Endogenous viruses are retained in the genome for long periods. For example, the Type C endogenous retrovirus of baboons is found in all species of Old World monkeys but not in New World monkeys. The sequence therefore invaded the germ line subsequent to the split between Old World monkeys and New World monkeys, and it remained within the Old World lineage through multiple speciation events.

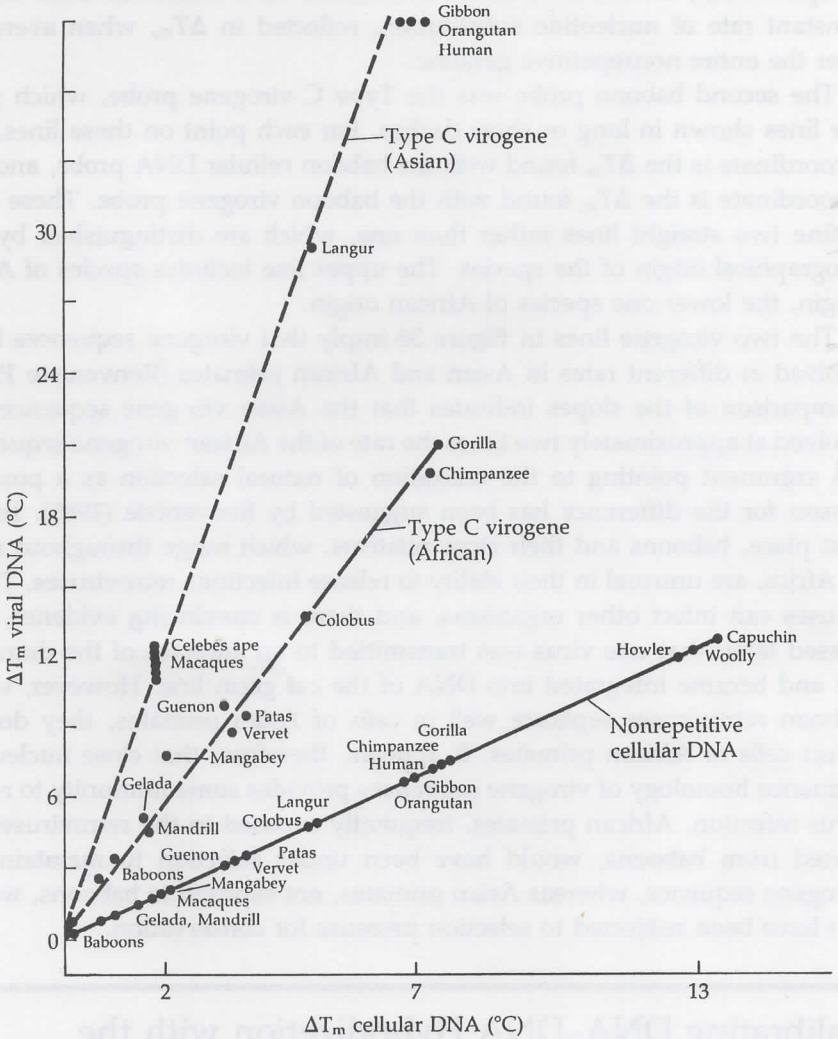


Figure 36. Evolutionary rate of baboon Type C endogenous retrovirus (dashed lines) as compared with nonrepetitive cellular DNA (solid). In African primates, these virogenes have evolved almost threefold faster than nonrepetitive cellular DNA, and in Asian primates they have evolved about sixfold faster. (After Benveniste 1985.)

The Type C endogenous retrovirus of baboons also illustrates the rapid rate of evolution characteristically found in virogenes. The evidence is shown in Figure 36, which summarizes data from DNA hybridization experiments. Figure 36 summarizes data obtained with two types of probe DNA from the baboon *Papio cynocephalus*, when hybridized with cellular DNA from other primate species. The first probe was baboon cellular DNA, which gave the solid line plotted with equal values on both axes. Note that the DNA from all species falls on the line, which is evidence for a molecular clock with a constant rate of nucleotide substitution, reflected in ΔT_m, when averaged over the entire nonrepetitive genome.

The second baboon probe was the Type C virogene probe, which gave the lines shown in long or short dashes. For each point on these lines, the x-coordinate is the ΔT_m found with the baboon cellular DNA probe, and the y-coordinate is the ΔT_m found with the baboon virogene probe. These data define two straight lines rather than one, which are distinguished by the geographical origin of the species. The upper line includes species of Asian origin, the lower one species of African origin.

The two virogene lines in Figure 36 imply that virogene sequences have evolved at different rates in Asian and African primates (Benveniste 1985). Comparison of the slopes indicates that the Asian virogene sequence has evolved at approximately two times the rate of the African virogene sequence. An argument pointing to the relaxation of natural selection as a possible reason for the difference has been suggested by Benveniste (1985). In the first place, baboons and their close relatives, which range throughout most of Africa, are unusual in their ability to release infectious retroviruses. These viruses can infect other organisms, and there is convincing evidence, discussed later, that one virus was transmitted to an ancestor of the domestic cat and became integrated into DNA of the cat germ line. However, while baboon retroviruses replicate well in cells of Asian primates, they do not infect cells of African primates. It appears, therefore, that close nucleotide sequence homology of virogene sequences provides some immunity to retrovirus infection. African primates, frequently exposed to the retroviruses released from baboons, would have been under selection to maintain the virogene sequence, whereas Asian primates, not exposed to baboons, would not have been subjected to selection pressure for conservation.

L Calibrating DNA–DNA Hybridization with the Molecular Clock

Given (1) that a ΔT_m of 1°C results from 1% nucleotide mismatches and (2) that a ΔT_m of 1°C occurs in 10 million years of evolution, what rate of nucleotide substitution per nucleotide site per year corresponds to a ΔT_m of 1°C? The slopes of the Asian and

African lines in Figure 36 are 6.0 and 2.8, respectively, relative to a value of 1.0 for nonrepetitive cellular DNA. What are the corresponding rates of evolution of Type C virogene sequences in terms of nucleotide substitution per site per year, using the rate just calculated for nonrepetitive cellular DNA? In Figure 36, the ΔT_m of the baboon–human comparison is 7.3°C. What is the predicted length of time of evolutionary divergence between these species?

ANSWER One percent nucleotide mismatches equals 0.01 mismatches per nucleotide site, a number sufficiently small that it need not be corrected for multiple substitutions. A value of ΔT_m of 1°C represents 0.01 mismatches per site divided by 10×10^6 years, or 1.0×10^{-9} substitutions per nucleotide per year. The Type C virogene rates are $2.8 \times 1.0 \times 10^{-9}$ substitutions per nucleotide per year for the African sequences. In the baboon–human comparison, a ΔT_m of 7.3°C would correspond to a divergence time of $7.3 \times (1/2) \times 10.0 \times 10^6 = 36.5 \times 10^6$ years. The factor of 1/2 enters because a total evolutionary time of 10 million years represents 5 million years since divergence (Figure 10).

It is of some interest in Figure 36 that the sequence of the human Type C virogene lies on the line characteristic of Asian primates. Taken at face value, this observation implies that a significant portion of human evolution may have occurred in an Asian setting (Benveniste 1985), which is contrary to the popular belief (and mounting genetic evidence) favoring Africa.

That endogenous retroviruses can undergo horizontal transfer between species that are only remotely related phylogenetically, and subsequently be incorporated into the germ line, is now well documented. Examples include the transfer and germ line integration of retroviral sequences from ancestors of modern primates to ancestors of carnivores, other primates, and birds; and the horizontal transfer from ancestors of modern rodents to ancestors of carnivores, primates, and artiodactyls (cleft-hoofed mammals).

Figure 37 provides evidence for the horizontal transfer of an endogenous retrovirus (named *RD*-114) from an ancestor of modern baboons (genus *Papio*) to an ancestor of the domestic cat (Benveniste 1985). The baboon-type virogene is not found in the lion, leopard, or cheetah, but it is in species closely related to the domestic cat. Thus, the baboon-type retrovirus must have been horizontally transferred after the divergence of the leopard cat but prior to the divergence of the black-footed cat. In terms of ΔT_m the bracketed values for the transfer are 1.6°C to 2.2°C. Assuming five million years of divergence per °C in ΔT_m (see Box L), the bracketed times for the transfer are 8–11 million years. In the phylogeny of baboons, the ΔT_m values demarcate the horizontal transfer as occurring after the divergence of the mandrill but prior to the divergence of the gelada.

By means of an as-yet poorly defined mechanism, retroviruses can occasionally incorporate genes of the host into their own genetic material. The

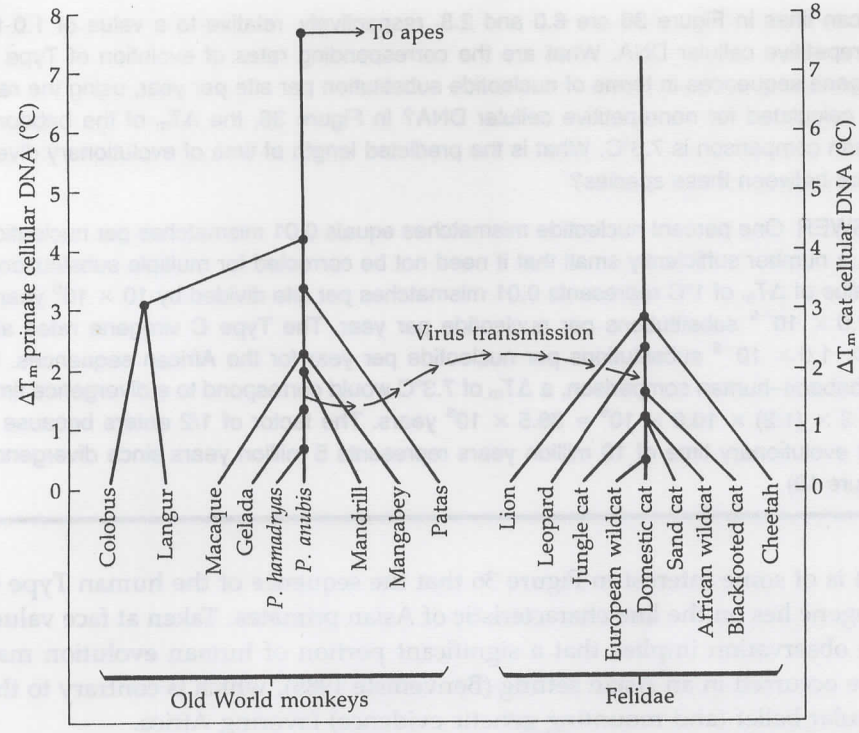


Figure 37. Domestic cats and their close relatives contain an endogenous retrovirus derived by horizontal transfer from ancestors of baboons. The time of the horizontal transfer can be determined by aligning the cat and baboon phylogenies according to the melting temperature of renatured nonrepetitive cellular DNA. In the cat phylogeny, the virus appears after the divergence of the leopard but before the divergence of the black-footed cat. In the baboon phylogeny this interval corresponds to the time between the divergence of the mandrill and the divergence of the gelada. In absolute time, the horizontal transfer appears to have occurred 8–11 million years ago. (After Benveniste 1985.)

best evidence on this point comes from the occurrence of oncogenes in certain retroviruses, which are associated with the cancerous growth of infected cells. Retroviral oncogenes, many of which are related to cellular growth factors, share unmistakable DNA sequence similarity with homologous genes found in the host organism and related species (Bishop 1981, 1982). Retroviruses can therefore bring about the horizontal transfer of genes between species. However, it remains to be determined whether other types of genes undergo horizontal transfer, by whatever mechanisms, at a rate that is of importance in the evolutionary process.

MITOCHONDRIAL AND CHLOROPLAST DNA EVOLUTION

In animals, mitochondria are usually inherited through the egg cytoplasm (maternal inheritance) and are genetically uniform within an individual. The mitochondrial genome consists of a single circular DNA molecule, denoted mtDNA, the size of which varies over a remarkably narrow range in different species (15.7–19.5 kb), averaging about 16 kb. (One bizarre exception to this is a deep-sea scallop whose mtDNA is 34 kb in length; see Snyder et al. 1987). Human mtDNA, illustrated in Figure 38, is fairly typical. It contains

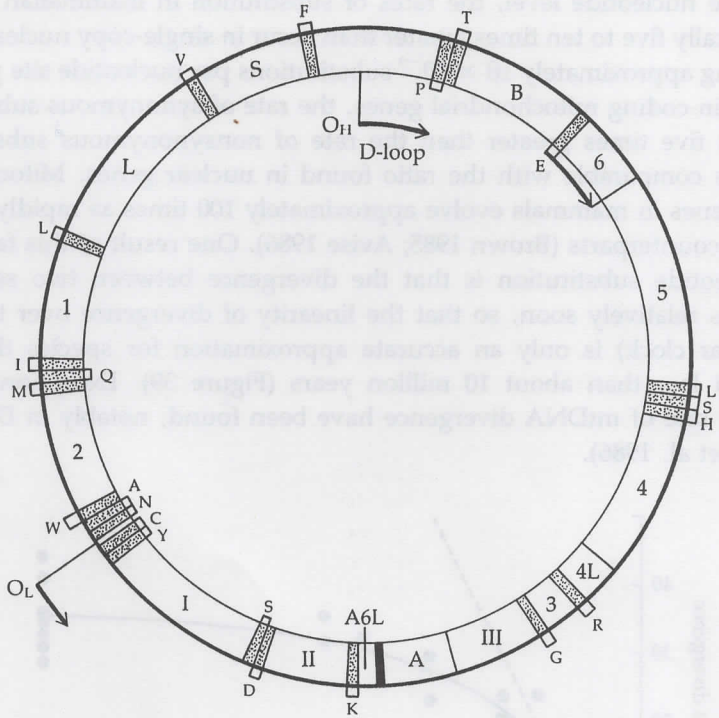


Figure 38. Organization of human mitochondrial DNA is typical of that in other mammals. It contains 13 protein-coding regions, 12 of which are essential in electron transport and ATP synthesis, plus genes for 22 tRNA molecules and two rRNA molecules. The typical length is 16,569 nucleotide pairs. In this diagram, the tRNA genes are indicated with stippling, L and S code for rRNA molecules, and the other labeled regions between the heavy and light strands code for proteins. The heavy and light lines represent the complementary strands of the mitochondrial DNA molecule. The arrows at O_H and O_L indicate the origin and direction of replication of each strand. H is the coding (template) strand for both rRNA genes and all proteins except number 6, and the coding strand for each tRNA gene is the strand nearest the label. (From Brown 1983.)

control regions for the initiation of DNA replication, plus genes for two rRNA molecules, 22 tRNA molecules, and 13 proteins — 12 of which code for subunits of enzyme complexes that carry out electron transport and ATP synthesis. The genetic code of mammalian mitochondria differs from the standard code (Table 3) in that ATA codes for Met, TGA codes for Trp, and AGR codes for End (termination of protein synthesis); thus, every codon in the mitochondrial code can be written as either *NNY* or *NNR*. Animal mitochondria also contain several hundred enzymes used in metabolic functions, but these are coded for by nuclear genes, and the enzymes are transported into the mitochondria.

At the nucleotide level, the rates of substitution in mammalian mtDNA are typically five to ten times greater than occur in single-copy nuclear genes, averaging approximately 10×10^{-9} substitutions per nucleotide site per year. In protein-coding mitochondrial genes, the rate of synonymous substitution is about five times greater than the rate of nonsynonymous substitution, which is comparable with the ratio found in nuclear genes. Mitochondrial tRNA genes in mammals evolve approximately 100 times as rapidly as their nuclear counterparts (Brown 1985; Avise 1986). One result of this faster rate of nucleotide substitution is that the divergence between two sequences saturates relatively soon, so that the linearity of divergence over time (the molecular clock) is only an accurate approximation for species that have diverged less than about 10 million years (Figure 39). Exceptions to the elevated rate of mtDNA divergence have been found, notably in *Drosophila* (Powell et al. 1986).

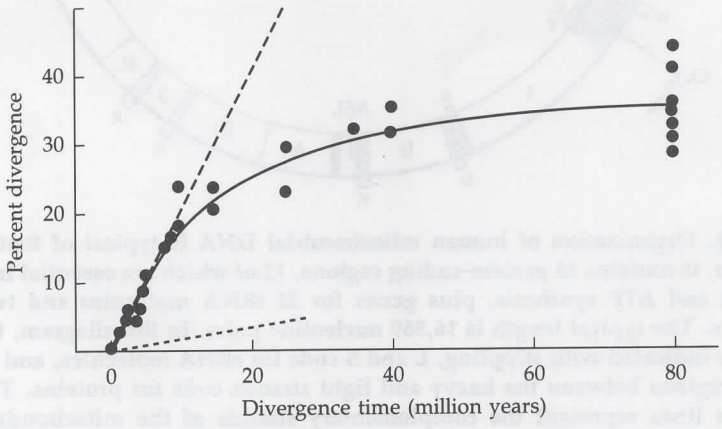


Figure 39. Relationship between the percentage of sequence divergence (100*d*) and divergence time. The points represent estimates from pairwise comparisons of restriction endonuclease cleavage maps. The initial rate of mtDNA sequence divergence is shown by the longer dashed line and the rate of divergence of single-copy nuclear DNA by the short dashed line (below). (From Brown et al. 1979.)

The extraordinary rate of substitution in mammalian mtDNA is thought to result from the occurrence of a high rate of mutation during mtDNA replication, either through a high rate of nucleotide misincorporation or a low efficiency of repair. Support for the latter view comes from the observation that the mitochondrial DNA polymerase lacks the proofreading function for error correction that occurs in the nuclear DNA polymerase.

Because mammalian mtDNA evolves so rapidly and tends to be highly polymorphic within species, it provides a useful tool with which to study the geographical structure of subpopulations within a species, or genetic differentiation between closely related species. For example, one restriction enzyme study of 87 pocket gophers (*Geomys pinetis*) yielded 23 distinct mtDNA types and revealed a strong geographical pattern in the occurrence of the types (Avise et al. 1979).

Human mtDNA Divergence

M

The mtDNA molecules of 21 humans of diverse geographic and racial origin were digested with 18 restriction enzymes, 11 of which exhibited one or more fragments in which size polymorphism occurred (Brown 1980). All restriction fragment length polymorphisms could be explained by single-nucleotide differences; thus there was no evidence for insertions, deletions, or other mtDNA rearrangements. Altogether, 868 nucleotide sites were assayed for differences among individuals, and the average number of differences per nucleotide site per individual was estimated at 0.0018. Assuming that mammalian DNA undergoes sequence divergence at the rate of $5–10 \times 10^{-9}$ nucleotide substitutions per site per year, and that the rate is uniform in time, calculate the length of time since all of the 21 contemporary mtDNA molecules last shared a common ancestor.

ANSWER With an average number of differences per nucleotide site per individual of 0.0018, and given an average rate of divergence of $5–10 \times 10^{-9}$ per site per year, the time of the last common ancestor would be between $0.0018/(10 \times 10^{-9})$ and $0.0018/(5 \times 10^{-9})$, or 180,000–360,000 years. That is, the present mtDNA genetic diversity in this sample of individuals could have originated from a single mitochondrion present in a female as recently as 9000 to 18,000 generations ago, assuming 20 years per generation. One possible explanation of this finding is that the human population might have undergone a bottleneck in population size at about that time. In the popular press, this conclusion has been interpreted as evidence for the existence of a unique female ancestor of all living humans, namely Eve. However, in the list of principles of neutral evolution discussed earlier in this chapter, Principle 4 says that among alleles destined to be fixed, the average time to fixation equals $4N_e$ generations. The corresponding time to fixation for a *mitochondrial* genome is N_e, because only females transmit mitochondria and mitochondria are genetically haploid. (Actually the rate of substitution for mitochondrial genes depends on other factors as well; see page 425.) Assuming that a single mitochondrial genome became fixed in 9000–18,000 generations, and

assuming that these values bracket the average, the corresponding range of estimates of $\langle N_e \rangle$ (for mitochondrial DNA) is 9000–18,000 individuals. This range appears reasonable for small societies of the sort that have characterized most of human evolution. Moreover, apart from the correction factor of 1/4, the average effective size of human populations in regard to mitochondrial DNA may not be representative of the average effective size in regard to nuclear DNA.

Cann et al. (1987) examined variation in the mtDNA of 147 humans in five populations using 12 restriction enzymes, sampling an average of 370 restriction sites per individual, or 9% of the mtDNA genome were assayed per individual. A total of 195 polymorphic sites were found in the genome, and the precise location on the mtDNA sequence of all polymorphic sites was identified. Altogether there were 133 different mtDNA types in the samples. When the mtDNA haplotypes were assembled into a phylogenetic tree, the common ancestor to all of the observed haplotypes (coalescence) occurred 140,000 to 290,000 years ago, assuming there is 2–4% divergence in mtDNA sequences per million years. There is some controversy over this figure (Saitou and Omoto 1987), centered largely on the rate of nucleotide divergence in human mtDNA. Cann et al. (1987) obtain the estimate of 2–4% per million years based on times of colonizations of ancient human populations, while Nei's (1985) figure of 0.71% per million years gives a coalescence time of 400,000 years.

Mammalian mtDNA is unique in undergoing a rapid rate of nucleotide substitutions but a slow rate of structural changes such as insertions and deletions. The DNA of other organelles shows different patterns of evolution, which are summarized qualitatively in Table 4 (Palmer 1985, 1988). All four possible patterns of fast and slow evolution are found.

Chloroplast DNA (cpDNA) ranges in size from 135 to 160 kb, and it occurs in multiple copies in each chloroplast. Its structural organization is conserved in higher plants, and the rate of synonymous nucleotide substitution is approximately 1×10^{-9} substitutions per site per year. Thus, the evolution of cpDNA is conservative in regard to both sequence and structure (Curtis and Clegg 1984). The rate of evolution of fungal mtDNA is at the other extreme — rapid in both sequence and structure.

The mtDNA of angiosperm plants has the opposite pattern of evolution as in animals. In sequence evolution the rate is slow, but in structural evolution the rate is fast. In plants, the mtDNA genome is large and highly complex. In some instances a single molecule can resolve itself into smaller circles and even linear molecules. For example, in the turnip *Brassica campestris*, a 218-kb molecule undergoes an internal recombination event that produces smaller circles of 135 kb and 85 kb. Maize mtDNA contains six pairs of repeated sequences that can undergo recombination and create a variety

Table 4. Rates of sequence and structural evolution in organelle DNA.

Genome	Rate of nucleotide substitution	Rate of structural evolution
Angiosperm cpDNA	Slow	Slow
Angiosperm mtDNA	Slow	Rapid
Mammalian mtDNA	Rapid	Slow
Fungal mtDNA	Rapid	Rapid

of structural derivatives. In addition, many plant mitochondria contain autonomously replicating plasmid DNA molecules, and they are also capable of incorporating segments of cpDNA. Why plant mtDNA genomes are so large, complex, and variable in size is not understood.

Population Genetic Theory for Cytoplasmically Transmitted Elements

The population genetics of organelles is unusual because of their (usually) uniparental transmission and because many copies are passed from the mother to the progeny through the egg. Uniparental transmission has important implications in the operation of natural selection, since it is an equivalent to a haploid clonal population structure, and pure selection models can only maintain polymorphism in such populations if the fitnesses are frequency dependent. From the outset, then, uniparental transmission makes it less likely for polymorphisms to be maintained by natural selection, even if epistatic effects with the nuclear genome are allowed (Clark 1984). The widespread polymorphisms observed in mtDNA must then be attributed largely to high mutation rates, just as the rapid substitution rate was attributed to high mutation rates. Polymorphisms can also be maintained by interspecific hybridization, and it is possible to obtain estimates of rates and directions of interspecific matings from nuclear and mtDNA data (Asmussen et al. 1987).

The theory of random genetic drift for organelles is more complex than that for nuclear genes because individual cells have many organelles that are divided among daughter cells, which means an additional level of sampling when heteroplasmic cells divide. (Heteroplasmic cells — cells having more than one mtDNA type — can arise by mutation or partial paternal transmission.) Models of the dual sampling process have been examined in some detail (Chapman et al. 1982; Birky et al. 1983; Takahata 1983, 1984). These

models require some level of heteroplasmy, and although early empirical studies did not detect heteroplasmy, it has now been described in crickets (Harrison et al. 1987), *Drosophila* (Hale and Singh 1986; Solignac et al. 1983, 1984, 1987), lizards (Densmore et al. 1985), mice (Boursot et al. 1987), cattle (Hauswirth and Laipis 1982), frogs (Monnerot et al. 1984), and treefrogs and bowfin fish (Bermingham et al. 1986). Heteroplasmy can be maintained by a steady-state balance between the forces of random genetic drift and mutation, but heteroplasmy is most frequently observed in fragment length polymorphisms, in which variants differ in number of copies of a small repeat. Simple deterministic models show that heteroplasmy can be stably maintained by infrequent paternal transmission (leakage), by natural selection, or by bidirectional mutation (such as the gain and loss events one would expect for changes in copy number of a small repeat; Clark 1988).

The argument that variation in organelle DNA is adaptively maintained would be more plausible if there were clear examples of associated phenotypic variation. For example, there are very clear examples of cytoplasmically transmitted phenotypic variation in plants, including leaf variegation and cytoplasmic male sterility (Grun 1976; Levings 1983). Similarly, cytoplasmically transmitted drug resistance genes have been shown to be associated with the mitochondrial genome of yeast. However, tests with *Drosophila* suggest that mtDNA variation contributes very little to overall variation in viability (Clark and Lyckegaard 1988). On the other hand, distributions of heteroplasmy in the field cricket are consistent with a model of mutation–selection balance, with smaller genomes favored by selection (Rand and Harrison 1986). The potential importance of mtDNA variation in human health was revealed in the implication of mitochondrial DNA defects in the muscle disease known as mitochondrial myopathy.

This chapter has summarized diverse aspects of the explosive growth in molecular population genetics that has accompanied successive improvements in the techniques of molecular biology and recombinant DNA. Molecular population genetics has become one of the most exciting and challenging growth areas of modern evolutionary biology. As DNA sequence data become ever more extensive and methods of computer data storage, retrieval, and analysis undergo further refinement, the role of population genetics in interpreting patterns of variation and in suggesting new types of analyses will assume augmented importance. These developments serve to emphasize the point that Crick (1979) made many years ago, that "a molecular biologist who wishes to discuss the evolution of the eukaryotic genome will need to know not only a lot about the way DNA and its transcripts can behave but also something about modern ideas on population genetics."

PROBLEMS

1 The following is a small portion of the gene coding for 6-phosphogluconate dehydrogenase in two natural isolates of *E. coli*. Each dot in the sequence represents 10 nucleotides. Infer the correct translational reading frame of the sequences and estimate

a. The number of amino acid differences/site;

b. The number of nucleotide differences/site;

c. The number of nonsynonymous substitutions per nonsynonymous site (regarding codon sites 1 and 2 as nonsynomyous);

d. The number of synonymous substitutions per synonymous site (regarding codon site 3 as synonymous).

```
1 CTGACCAAAATCGCCGCCGTAGCTGAAGACGGTGAACCATGCGTTACCTATATTGGTGCC
2 ----AGC-G-----G--G--T--------------G--G--T--G--T-----A-----T
```

2 In the human immunodeficiency virus HIV, which causes acquired immune deficiency syndrome (AIDS), the rate of nucleotide evolution has been estimated at about 0.01 substitutions per synonymous site per year. Two viruses isolated in 1983 in Zaire and San Francisco differ in approximately 1/3 of their synonymous sites. Estimate the year in which the viruses last shared a common ancestor. (Data from Li et al. 1988.)

3 The data below give the proportion of nucleotide sites that differ in a gene in four RNA viruses (Yokoyama et al. 1988). HIV1 and HIV2 are two rather distinct types of human immunodeficiency viruses, VISNA is a lentivirus, and MMLV is a mouse cancer-causing virus. Estimate the number of nucleotide substitutions per site using these data. What do the numbers imply about the evolutionary relationships among the viruses?

	HIV2	VISNA	MMLV
HIV1	0.34	0.54	0.62
HIV2		0.52	0.63
VISNA			0.63

4 What inference would you make regarding the selective constraints on a region of DNA in which the rate of evolution was 5×10^{-9} nucleotide substitutions per site per year?

5 What might you infer about the evolutionary forces affecting a coding region in which the rate of amino acid replacement was greater than the rate of synonymous nucleotide substitution?

6 Ribosomal RNA forms a complex secondary structure in which many regions of the molecule are folded back and undergo base pairing with

complementary nucleotide sequences elsewhere in the same molecule. What pattern of nucleotide sequence evolution might be expected in these paired regions?

7 What is the largest value of D that makes sense in Equation 7.21 and what does it mean?

8 If the rate of nucleotide evolution along a lineage is 0.5% per million years, what is the rate of substitution per nucleotide per year? What is the total rate of divergence of two lineages?

9 For the model of bacterial insertion sequences leading to Equation 7.26, what are the expected proportions of strains with i elements when $\alpha = 1/4$ and $\theta = 1/4$? When $\alpha = 1/4$ and $\theta = 3/4$?

10 Insertion sequence IS*1* in *E. coli* has a distribution of copy number that can be fitted using Equation 7.26 with $\alpha = 1/5$ and $\theta = 5/6$. Among 71 strains, the number with i copies of IS*1* was 11, 14, 8, 6, 7, and 25 for $i = 0, 1, 2, 3, 4$, and ≥ 5. What are the expected numbers and the χ^2 value?

11 Comparison of nonrepetitive cellular DNA of human and orangutan gives a ΔT_m value of 3.6°C. Assuming that these species diverged 13 million years ago, what is the rate of evolution of ΔT_m? Assuming that a ΔT_m of 1°C corresponds to 1% nucleotide mismatches, what is the rate of nucleotide substitution?

12 If the effective size of an ideal diploid population is N with respect to autosomal genes, what is it with respect to

a. X-linked genes?

b. Y-linked genes?

c. mtDNA?

13 In comparing the mitochondrial DNAs of two Japanese monkeys in the species *Macaca fuscata* using 17 six-base restriction enzymes, Hayasaka et al. (1988) observed 47 restriction fragments in individual J-1, 42 in individual J-2, and 39 shared by both. How many nucleotides were sampled in this comparison, and what is the estimated proportion of polymorphic nucleotide sites?

14 Considering Equations 7.10 and 7.11 for evolutionary divergence d in terms of probability of a restriction site change per unit time M, what value of M for six-base recognition sites corresponds to a d of 1%? What proportion of shared restriction fragments does this correspond to?

15 Restriction mapping of mitochondrial DNAs in the rhesus monkey *Macaca mulatta* and the crab-eating monkey *M. fascicularis*, using 17 six-base restriction enzymes, gave 43 and 49 restriction fragments in the species, including 33 shared fragments (Hayasaka et al. 1988). Use these data to estimate the expected number $\langle d \rangle$ of nucleotide substitutions per site in the mitochondrial DNA of these species.

16 For an ideal population of effective size N, the average time to loss of a new mutation destined to be lost is $2\ln(2N)$, and the average time to fixation of a new mutation destined to be fixed is $4N$. For what values of N does

a. Fixation time = 10 × loss time?

b. Fixation time = 100 × loss time? (Hint: Solve by iteration, as in Equation 7.10.)

17 For the model of gene conversion with gene identities given in Equations 7.24a and b, what value of λ makes the organization of the gene family irrelevant in the sense that $f = c_1 = c_2$? What is the common value in this case? (λ is the probability that a particular member of the gene family becomes converted in any one generation.)

18 For the model of gene conversion with gene identities given in Equations 7.24a and b, what are the values of f and $c_1 = c_2$ when $\lambda = \mu$? (The equations assume $4N\mu << 1$.)

19 In a repetitive gene family being eliminated from the genome by deletion, if the fraction of sequences present at time 0 that are still present at time t equals $\exp(-ht)$, show that the half-life of the sequences equals $-\ln(1/2)/h$.

20 For a repetitive gene family eliminated as described in Problem 19, show that the average persistence of an element is $1/h$.

Chapter Eight

Evolutionary Genetics of Quantitative Characters

Darwin formulated his ideas about evolution by natural selection based on observations of phenotypic variation. Even today, many important problems in evolutionary biology begin with observations of phenotypic variation. The variable traits often depend on the influence of many genes: for example, traits such as lipid storage, body size, flight power, or number of inflorescences. Simple crossing experiments demonstrate that these traits are not determined by single genes because the offspring do not fall into discrete classes with tidy Mendelian ratios. Instead, what is observed is a general resemblance between parents and offspring, suggesting that there is an underlying genetic basis to the trait, but that the genetic transmission is complex.

A wealth of powerful statistical tools have been developed for analyzing such **polygenic traits** that do not show simple Mendelian transmission. These approaches allow not only description of the genetic basis of observed phenotypic distributions, but they also provide means of predicting the distributions of phenotypes among offspring from the parents' phenotypes. Most polygenic traits are influenced by the environment to varying degrees, and they are often called **multifactorial traits** to emphasize their determination by multiple genetic and environmental factors. For example, variation in human weight is partly due to genetic differences among individuals and partly due to environmental factors such as exercise and nutrition. The study of polygenic inheritance goes beyond the oversimplified nature-versus-nurture dichotomy, because it is concerned with specifying, in precise quantitative terms, the relative importance of nature, nurture, and their interactions in accounting for variation in phenotype among individuals. Another com-

pelling reason to study polygenic inheritance is that natural selection occurs largely at the level of the phenotype, i.e., *fitness is a multifactorial trait*. Since natural selection operates on phenotypes, there arises an immediate problem in understanding the consequences of phenotypic evolution on changes that occur at the molecular level, and we are still a long way from understanding the molecular basis of complex phenotypes. A great challenge facing population geneticists is to unify the principles of molecular evolution with those governing evolution at the phenotypic level.

TYPES OF QUANTITATIVE TRAITS

Multifactorial traits may be considered as resulting from the combined effects of many quantities, some genetic in origin and some environmental, and for this reason they are often called **quantitative traits**. The study of quantitative traits constitutes **quantitative genetics**.

Three types of quantitative traits may be distinguished:

1. Traits for which there is a continuum of possible phenotypes are **continuous traits**, and examples include height, weight, milk yield, and growth rate. The distinguishing feature of continuous traits is that the phenotype can take on any one of a continuous range of values. In theory, there are infinitely many possible phenotypes, among which discrimination is limited only by the precision of the instrument used for measurement. However, in practice, similar phenotypes are often grouped together for purposes of analysis.
2. Traits for which the phenotype is expressed in discrete, integral classes are **meristic traits**, and examples include number of offspring or litter size, number of ears on a stalk of corn, number of petals on a flower, and number of bristles on a fruit fly. The distinguishing feature of meristic traits is that the phenotype of an individual is given by a simple integer which equals the number of elements of the trait that the individual displays. For example, a popular meristic trait used in experimental studies of quantitative genetics in *Drosophila* is the number of bristles that occur on the abdominal sternites, typically the sixth in females and the fifth in males (Figure 1). Normally there are 14 to 24 bristles per sternite. A male with 19 bristles on the fifth abdominal sternite therefore has a phenotype of 19. When the number of possible phenotypes of a meristic trait is large (as it is with abdominal bristle number) then the line between continuous traits and meristic traits becomes indistinct. The distribution of numbers of abdominal bristles in a sample of *D. melanogaster* appears in Figure 2.

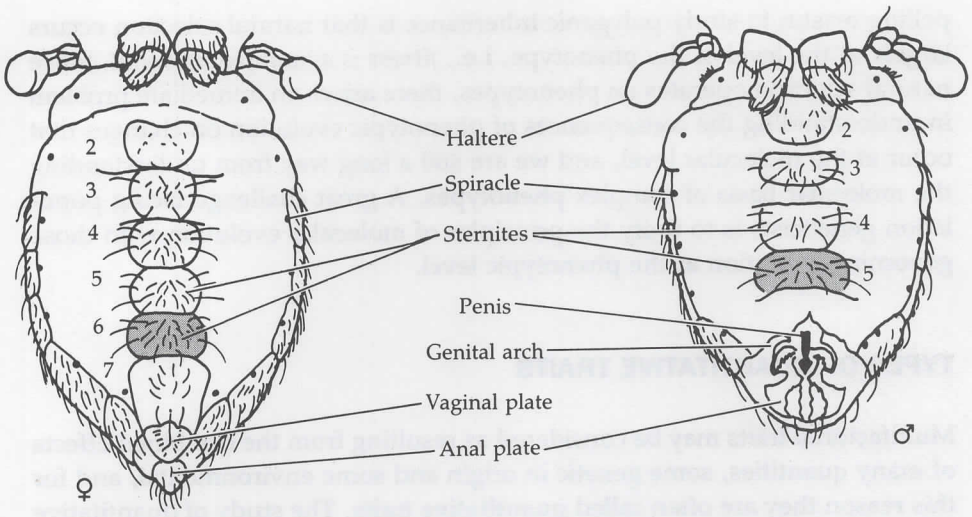


Figure 1. Ventral abdomen of females and males of *Drosophila melanogaster*, indicating the sternites. The number of sternital bristles (also called abdominal bristles) provides a convenient trait for experimental studies in quantitative genetics. The enumerated bristles are usually those on the fifth or sixth (shaded) sternite in females (or their sum), and the fourth or fifth (shaded) sternite in males (or their sum).

3. The third category of quantitative traits consists of **discrete traits**, which are either present or absent in any one individual. In these cases, the multiple genetic and environmental factors combine to determine an underlying risk or **liability** toward the trait. Liability values are not directly observable. However, individuals that actually express the trait are assumed to be ones whose liability values are greater than some threshold or triggering level. Traits of this type are **threshold traits**, and examples in human genetics include diabetes and schizophrenia. With threshold traits, studies of affected individuals and their relatives permit inferences to be made about underlying values of liability. These methods are discussed later in this chapter.

Quantitative traits are of utmost importance to plant and animal breeders, since agriculturally important characteristics such as yield of grain, egg production, milk production, efficiency of food utilization by domesticated animals, and meat quality are all quantitative traits. Even as modern methods of genetic engineering are applied to animal and plant improvement, quantitative genetics continues to play an important role because commercially desirable traits result from complex interactions among many genes. In ad-

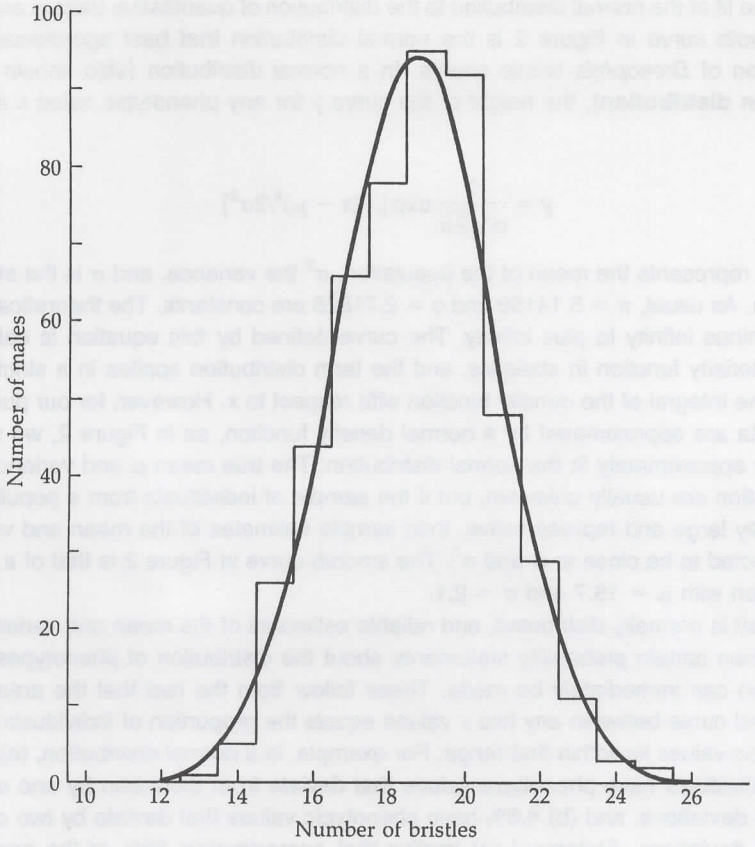


Figure 2. Number of bristles on the fifth abdominal sternite in males of a strain of *Drosophila melanogaster*. The smooth curve is that of a normal distribution with mean 18.7 and standard deviation 2.1. (Data from Mackay 1984.)

dition to being essential ingredients in plant and animal improvement programs, the principles of quantitative genetics, appropriately modified and interpreted, can be applied to the analysis of quantitative traits in humans and natural populations of plants and animals.

The Normal Distribution and the Central Limit Theorem

A

Many quantitative traits have a distribution of phenotypes that can be specified completely by the mean phenotype and the variance in phenotype. In these cases, the mean and variance determine the shape of the entire distribution. An example of such a two-parameter distribution is the familiar, bell-shaped normal distribution. In many

cases, the fit of the normal distribution to the distribution of quantitative traits is excellent. The smooth curve in Figure 2 is the normal distribution that best approximates the distribution of *Drosophila* bristle counts. In a normal distribution (also known as the **Gaussian distribution**), the height of the curve y for any phenotypic value x is given by

$$y = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-(x - \mu)^2/2\sigma^2\right]$$

where μ represents the mean of the population, σ^2 the variance, and σ is the standard deviation. As usual, $\pi = 3.14159$ and $e = 2.71828$ are constants. The theoretical range of x is minus infinity to plus infinity. The curve defined by this equation is called the normal density function in statistics, and the term distribution applies in a strict sense only to the integral of the density function with respect to x. However, for our purposes, if the data are approximated by a normal density function, as in Figure 2, we will say that they approximately fit the normal distribution. The true mean μ and variance σ^2 of a population are usually unknown, but if the sample of individuals from a population is sufficiently large and representative, then sample estimates of the mean and variance are expected to be close to μ and σ^2. The smooth curve in Figure 2 is that of a normal distribution with $\mu = 18.7$ and $\sigma = 2.1$.

If a trait is normally distributed, and reliable estimates of the mean and variance are known, then certain probability statements about the distribution of phenotypes in the population can immediately be made. These follow from the fact that the area under the normal curve between any two x values equals the proportion of individuals whose phenotypic values lie within that range. For example, in a normal distribution, (a) 31.7% of the individuals have phenotypic values that deviate from the mean by one or more standard deviations, and (b) 4.6% have phenotypic values that deviate by two or more standard deviations. Statement (a) implies that approximately 68% of the population have phenotypic values within one standard deviation from the mean. Similarly, statement (b) implies that approximately 95% of the population have phenotypic values within two standard deviations from the mean. The frequent occurrence of the normal distribution in practice results in part from the fact that the distribution of any quantity computed as the sum of many random, independent quantities converges to a normal distribution. Stated in terms of mathematical statistics, the distribution of any random variable equal to the sum of many independent, identically distributed random variables converges to a normal density as the number of random variables in the sum increases. This convergence constitutes the **central limit theorem**, which is one of the most important theorems in statistics. Many measurable quantities in the real world are determined by such sums of independent causes. For example, the multiple genetic and environmental factors that determine quantitative traits may be approximately additive in their effects, so the central limit theorem is expected to hold. For many characters, the factors appear to multiply in their effects, and in these cases a logarithmic transformation gives a better approximation to the normal distribution. In practice, such transformations are done based on the simple empirical observation that a better fit to normality can be achieved after transforming the data. The normal distribution's prevalence and its convenient mathematical characteristics make it very important in quan-

titative genetics. As illustrated in Figure 2, the normal distribution is of value not only when applied to continuous traits but also when applied to meristic traits when the number of possible phenotypes is relatively large. The key factor in arriving at a normal distribution is the independence of the component causal factors. Interdependence of causal factors does occur in quantitative genetics, and this can result in departure from the normal distribution.

RESEMBLANCE BETWEEN RELATIVES AND THE CONCEPT OF HERITABILITY

Everyone has, to one degree or another, a sort of street-sense about quantitative genetics. When we are introduced to a friend's brother or sister, we subliminally look for similarities and differences. We celebrate the birth of a child by claiming that it has its mother's nose or its grandfather's ears. When confronted by a pair of identical twins, we are often stunned by their similarities. We all know that there are many degrees of relationship, and it is reasonable to assume that closer familial relationships generally lead to greater phenotypic similarity. This simple concept, dressed up with a wealth of powerful statistical tools, is the essence of quantitative genetics.

The relationship that has received the most attention by quantitative geneticists is that of parent and offspring. The essential feature of the transmission of traits in Darwinian evolution is that offspring resemble their parents, and hence variability across generations is maintained. Indeed, even before the rediscovery of Mendel, Francis Galton was collecting detailed statistical data on resemblance of parents and offspring, and he devised a method known as regression for describing the statistical relationship. In this chapter we will first develop the central ideas of transmission of quantitative traits, and then we will develop models of Mendelian inheritance that account for many features of the transmission. Calculation of the degree of resemblance among relatives in terms of underlying Mendelian genetics was first provided by Fisher (1918), in a paper of great historical importance to population genetics, because it provided the first demonstration that multiple Mendelian genes could account for the observed patterns of transmission of multifactorial traits.

Figure 3 shows a plot of the mean of male offspring for a quantitative trait (y values) against the phenotypic value of the father (x values), displayed in a Galtonian manner. The line is the best fitting straight line, called the **regression line**, of offspring on parent. Regression is relevant to one of the primary aims in animal and plant breeding, which is to be able to improve attributes of the stock. An essential part of genetic improvement is to be able to predict what sort of offspring would be obtained from a given pair of

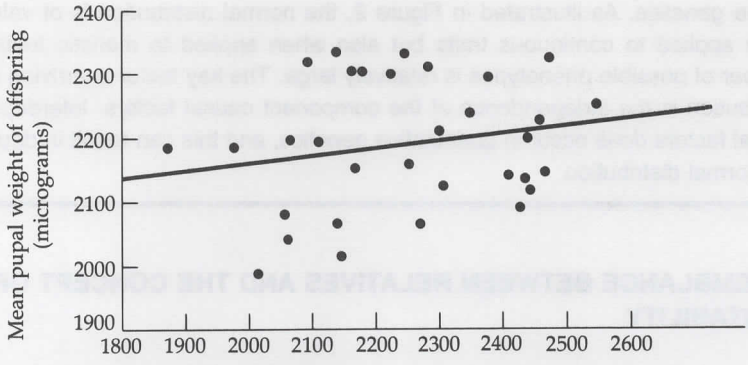


Figure 3. Mean weight of male pupae of the flour beetle *Tribolium castaneum*, plotted against pupal weight of their father (sire). Each point is the mean of about eight male offspring. The regression coefficient of male offspring weight on sire's weight is $\langle b \rangle = 0.11$, and h^2 is estimated as $2b$. (Courtesy of F. D. Enfield.)

parents. For quantitative traits, prediction cannot be done exactly, but a statistical description of the most likely offspring can be obtained by the procedure of plotting the parent-offspring regression. For reasons that will become clear in a moment, we are interested in the slope of the regression line. The slope is most easily expressed in terms of the covariance of x and y, as defined in Box B.

B Statistical Tools Used in Quantitative Genetics

For those lacking prior experience with statistical operations used in biology, a summary of the principal statistical quantities used in quantitative genetics might be helpful. Suppose we have n pairs of measurements (x_1,y_1), (x_2, y_2), . . . , (x_n, y_n); in an actual example, x_i might be hat size and y_i might be shoe size; or x_i might be the maternal milk yield and y_i the daughter's milk yield. As elsewhere, the symbol Σ below denotes summation and a bar over a variable denotes the sample mean. Thus $\bar{x} = (\Sigma x)/n$ is the sample mean of x; and the mean of the squares is $\overline{x^2} = (\Sigma x^2)/n$; and the mean of the products of x and y is $\overline{xy} = \Sigma xy/n$. The principal quantities that are used are:

1. $\bar{x} = (\Sigma x)/n$.
2. $\bar{y} = (\Sigma y)/n$.
3. $\text{Var}(x) = [\Sigma(x - \bar{x})^2]/n$. Note that $\text{Var}(x)$ is the sample variance of x, and that it can also be expressed as $\overline{x^2} - \bar{x}^2$. This estimate for the variance is actually a biased estimate, because it is slightly off the best estimate. An unbiased estimate of σ^2 can be obtained using the formula $\text{Var}(x) = [\Sigma x^2 - (\Sigma x)/n]/(n - 1)$.
4. $\text{Cov}(x,y) = [\Sigma(x - \bar{x})(y - \bar{y})/n = (\overline{xy}) - (\bar{x})(\bar{y})$. This quantity is the sample **covariance** of x and y.

5. $r = \text{Cov}(x,y)/\sqrt{\text{Var}(x)\text{Var}(y)}$ is the sample **correlation coefficient**.
6. $b = \text{Cov}(x,y)/\text{Var}(x)$ is the **regression coefficient**, which equals the slope of the regression line of y on x.

An important concept in statistics is the distinction between **parameters** and **estimators**. Descriptors that are calculated from a set of data to describe a sample (such as the sample mean and sample variance) are considered as estimates of the parameters that determine the true distribution. The sample is thought of as having been drawn from some perfect distribution (whose parameters we can never know), and the sample statistics give us a best guess at what that true distribution is. Parameters are generally distinguished from estimators by Greek symbols: μ is the parametric mean, σ^2 is the variance, σ_{xy} is the covariance of x and y, and ρ is the correlation. Another notation that we have found convenient throughout this book is, for example, $\langle\mu_x\rangle$ to denote the sample mean of x. With this convention, $\langle\mu_x\rangle = \bar{x}$, $\langle\sigma_x^2\rangle = \text{Var}(x)$, and $\langle\sigma_{xy}\rangle = \text{Cov}(x,y)$ are all sample estimates. When describing models of quantitative genetics, it is the true distributions that are of interest, and so the parameters are used. When describing the results of an experiment, it is more appropriate to use the notation for estimates.

The covariance and the correlation coefficient are convenient measures of the degree of association between x and y. If x and y are independent, then σ_{xy} and ρ are both zero. Since the covariance between any two variables measures their degree of association, the covariance may be positive or negative. Positive covariance means that values of x and y tend to increase or decrease together; negative covariance means that, as one variable increases, the other tends to decrease. The limiting values of the covariance are $-\sigma_x\sigma_y$ on the negative side, and $\sigma_x\sigma_y$ on the positive. The limits are achieved only when the variables demonstrate a perfect linear relationship with each other.

Returning now to Figure 3, if $\text{Cov}(x,y)$ represents the covariance between phenotypic values of fathers (sires) and those of their male offspring, and $\text{Var}(x)$ represents the variance of phenotypic values of the fathers, then the slope of the regression line is simply the regression coefficient, $\text{Cov}(x,y)/\text{Var}(x)$, which can be seen as follows. Suppose that the equation of the line is represented as

$$y = c + bx \tag{8.1}$$

where c and b are constants, and b is the slope. Taking means of both sides yields

$$\bar{y} = c + b\bar{x} \tag{8.2}$$

Subtracting the second equation from the first yields

$$y - \bar{y} = (c + bx) - (c - b\bar{x}) = b(x - \bar{x}) \tag{8.3}$$

Now multiply through by $x - \bar{x}$ to obtain

$$(x - \bar{x})(y - \bar{y}) = b(x - \bar{x})^2 \tag{8.4}$$

Taking means of both sides produces

$$\text{Cov}(x,y) = b\text{Var}(x) \tag{8.5}$$

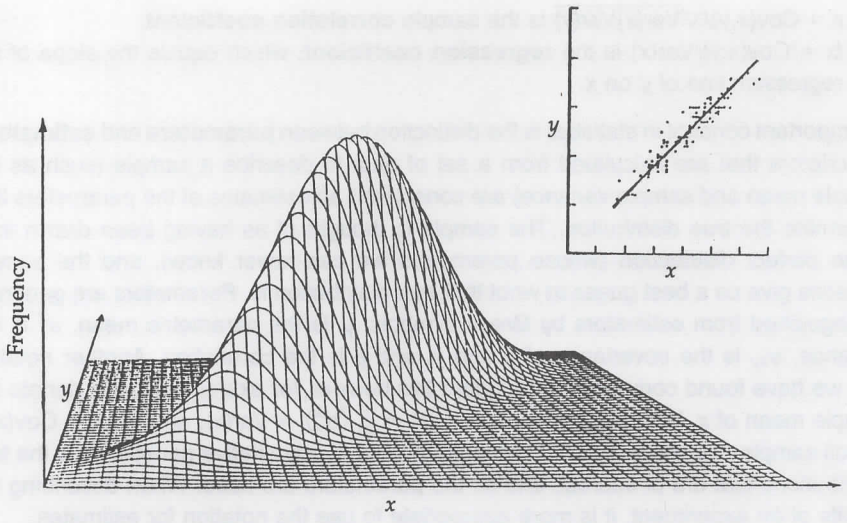


Figure 4. Bivariate normal distribution of correlated variables x and y, showing the ridge on the surface which results from the correlation. The inset shows pairs of (x,y) points sampled at random from the distribution and the resulting regression line. (After Sokal and Rohlf 1969.)

In other words, the slope b of the regression line equals

$$b = \text{Cov}(x,y)/\text{Var}(x) \qquad \textbf{(8.6)}$$

As noted, the slope is called the regression coefficient of offspring on one parent.

A geometrical interpretation of regression is illustrated in Figure 4, which shows the distribution, in two dimensions, of the variables x and y. The variables may represent, for example, the phenotypic values of parents (x) and offspring (y). If there were no association between x and y, the surface would be completely symmetrical and resemble an inverted bowl. The ridge in Figure 4 results from the association. When pairs of (x, y) values are sampled from the two-dimensional distribution, a plot of the points yields the sort of scatter diagram shown in the inset in Figure 4, in which the straight line is the regression line of y on x. The slope of the regression line is given by Equation 8.6.

Two extreme examples may help clarify parent–offspring regression. At one extreme, if there were no genetic contribution to the trait, then the scattergram as in Figure 4 might appear as a random scatter with no tendency to follow a line. In such a case, knowing the phenotype of the parents would not help to predict that of the offspring, because there would be no parent–offspring resemblance. On the other hand, even with no genetic variation, the points might nevertheless show a substantial tendency to follow a line. To see why this is so, consider families living in different environments. In

rich environments with plenty of food and resources, parents and offspring might all be big and strong, while in poor environments, parents and offspring might be small and sickly. A parent–offspring plot would show that big strong parents have big strong offspring, while small sickly parents have small sickly offspring—even though there is absolutely no genetic basis for the traits. The tendency of points to follow a line in a parent–offspring scattergram tells us nothing about the genetic basis for the trait, unless we are willing to make some claims (which hopefully have been tested experimentally) about the **environmental covariance** (the tendency of parents and offspring to resemble one another due to shared environments). Only if there is no environmental covariance will the parent–offspring regression indicate a degree of genetic influence on the resemblance. This concept is absolutely critical in human quantitative genetics, where the influence of shared environments can be subtle but strong.

Assuming now that the environmental covariance is zero, the regression coefficient b of offspring on one parent can be calculated for any random-mating population, and it indicates the degree to which the variance in the trait is determined by genetic variation. It is for this reason that the regression coefficient is related to an important quantity in quantitative genetics called **heritability**. There are two types of heritability that will be distinguished shortly, but for now, we note that the "narrow-sense" heritability (h^2) can be estimated from the relationship

$$b = h^2/2 \qquad \textbf{(8.7)}$$

The 1/2 occurs in Equation 8.7 because the regression involves only a single parent (the father, in the case of Figure 3), and only half of the genes from any one parent are passed on to the offspring. In Figure 3, $\langle b \rangle = 0.11$, so $\langle h^2 \rangle = 0.22$. Notice the considerable scatter among the points in the figure, which represents data from 32 families. Because this sort of scatter is typical, heritability estimates tend to be quite imprecise unless based on data from several hundred families. Note, however, that even with an enormous sample, there would be no less scatter to the points—we would just have a very accurate measure of how much scatter there is. One further point about Figure 3: in organisms such as mammals, the regression is better performed on the father's phenotype, rather than on the mother's, in order to avoid potential bias in the estimate of heritability caused by such maternal effects as uterine environment. In organisms where nurturing does not impart significant maternal effects, scattergrams can be constructed with the x axis being the average of the two parents (the **midparent**), and the y axis the offspring phenotypes. From this sort of plot, the regression coefficient is equal to the heritability (i.e., $b = h^2$ when the x axis is the midparent).

C Calculation of Heritability from Parent–Offspring Regression

This example of calculating h^2 from parent-offspring regression uses data from Cook (1965), who studied shell breadth in 119 sibships of the snail *Arianta arbustorum*. For computational convenience, the data have been grouped into six categories. We wish to estimate the heritability of shell breadth from these data.

Number of sibships	Midparent value (mm)	Offspring mean (mm)
22	16.25	17.73
31	18.75	19.15
48	21.25	20.73
11	23.75	22.84
4	26.25	23.75
3	28.75	25.42

Letting x refer to the midparent value and y refer to the offspring mean, then, $\bar{x} = 20.2626$, $\bar{y} = 20.1786$, $\Sigma x_i^2 = 49{,}823.4375$, $\Sigma y_i^2 = 49{,}267.1875$, $\langle \sigma_{xy} \rangle = 5.1826$, $\langle \sigma_x^2 \rangle = 8.1801$, and $\langle b \rangle = \langle h^2 \rangle = 0.63$. (In actual practice we might not want to group the data into categories, because there is some loss of accuracy from grouping. The regression coefficient for the ungrouped data is $b = 0.70$. In addition, it should be noted that there is substantial assortative mating for shell breadth, which makes the heritability estimate artificially large).

To this point we have shown that heritability can be used to measure the degree of resemblance between parents and offspring. Although the definition of heritability in terms of the regression coefficient between midparents and offspring is reasonable, at this level heritability is merely a descriptive, empirical quantity because it makes no assumptions about genetics. However, as soon as we postulate that there are Mendelian genes underlying the phenotypes, the calculations become a bit more complicated. We must add a genetic underpinning to the discussion—otherwise all that we could do is describe statistical relations among individuals. Once we bring Mendelian genetics into the picture, we will see that after a few critical assumptions are made, response to artificial selection of any kind is determined by the magnitude of the heritability. This is because the genetic basis of response to artificial selection comes from changes in gene frequencies and perhaps changes in linkage disequilibrium. Later we show that phenotypic variance is composed of components that can be estimated by means of appropriate experiments, and that some of the components of variation are related to the narrow-sense heritability. The correspondence between narrow-sense heritability and the components of variance will be derived theoretically in the

section on components of phenotypic variability, where we show that the definitions based on parent-offspring regression and response to selection are the same if the environmental covariance is zero.

ARTIFICIAL SELECTION AND REALIZED HERITABILITY

The deliberate choice of a select group of individuals to be used for breeding constitutes **artificial selection**. The most common type of artificial selection is **directional selection**, in which phenotypically superior animals or plants are chosen for breeding. Although artificial selection has been practiced successfully for thousands of years (for example, in the body size of domesticated dogs), only during this century have the genetic principles underlying its successes become clear. Understanding the genetic principles of artificial selection permits prediction of the rapidity and amount by which a population can be altered through artificial selection in any particular generation or small number of generations. The theory of artificial selection is also strongly motivated by the idea that natural selection may operate in a similar way. For example, if only those individuals with greater than a certain amount of body fat survive, or only those individuals with less than a critical rate of evaporative water loss survive, then natural selection acts on the distribution of phenotypes in much the same way that breeders select characters of agricultural importance.

Artificial selection in outcrossing, genetically heterogeneous populations is usually successful in that the mean phenotype of the population changes over generations in the direction of selection (provided the population has not previously been subjected to long-term artificial selection for the trait in question). In experimental animals, the mean of almost any quantitative trait can be altered in whatever direction desired by artificial selection. For example, in *Drosophila*, body size, wing size, bristle number, growth rate, egg production, insecticide resistance, and many other traits can be increased or decreased by selection. In domesticated animals and plants, birth weight, growth rate, milk production, egg production, grain yield, and countless other traits respond to selection. Figure 5 shows the results of a long-term selection program involving oil content in corn.

The general success of artificial selection in outcrossing species indicates that a wealth of genetic variation affecting quantitative traits exists. On the other hand, the mean phenotype of a population cannot usually be changed through artificial selection in genetically uniform populations, because genetic variation is required for progress under artificial selection. For example, in experiments with the Princess bean, Johannsen (1909) found that artificial

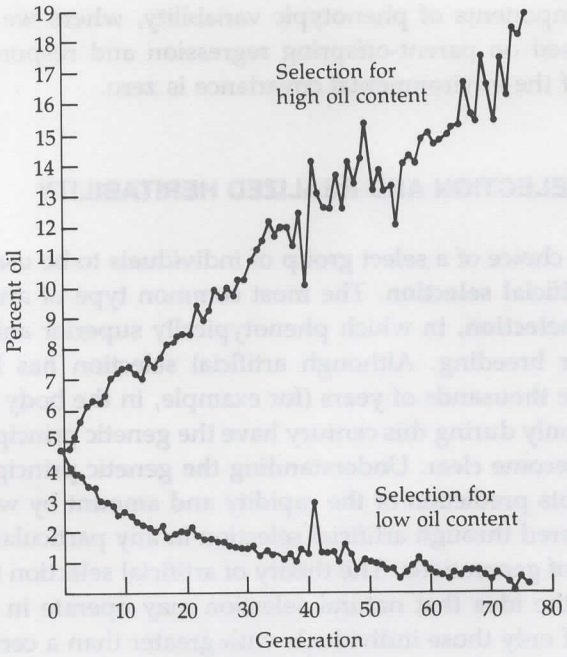


Figure 5. Results of a famous long-term experiment selecting for high and low oil content in corn seeds. Begun in 1896, the experiment has the longest duration of any on record, and still continues at the University of Illinois. Note the steady, linear rise in oil content shown by the upper curve. The lower curve started on a roughly linear path and continued so for about ten generations, but then the response tapered off, presumably because zero percent oil is an absolute lower limit for the trait. (After Dudley 1977.)

selection consistently resulted in failure when practiced within essentially homozygous lines. He obtained this result because, in genetically homozygous populations, the only source of genetic variation comes from new mutations. In contrast, since genetically variable populations usually respond to artificial selection, and genetically uniform populations do not respond, the response to artificial selection might be used as a measure of the extent of genetic variation in the trait. This notion of response reflecting variation will be formalized in the next section.

Prediction Equation for Individual Selection

When individuals are selected for breeding based solely on their own individual phenotypic values, the type of artificial selection is called **individual selection**. Figure 6 illustrates a variety of individual selection called **truncation selection**. The curve in panel (a) represents the normal distribution of a

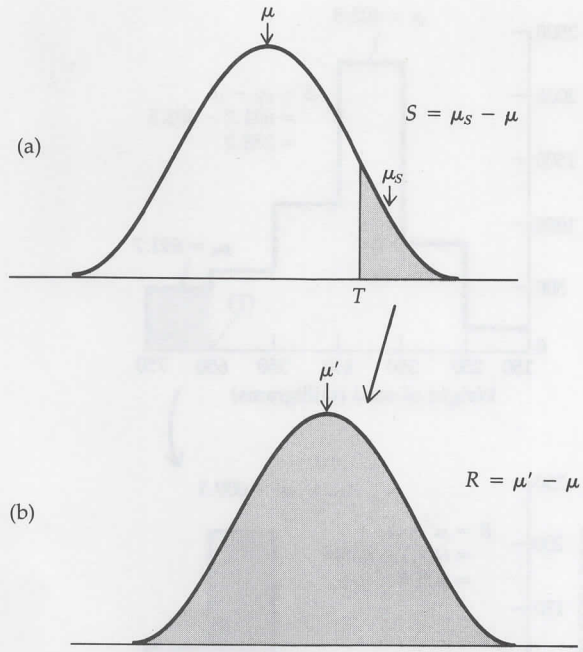


Figure 6. Diagram of truncation selection. (a) Distribution of phenotypes in parental population, with mean μ. Individuals with phenotypes above the truncation point (*T*) are saved for breeding in the next generation. The selected parents are denoted by the shading and their mean phenotype by μ_s. (b) Distribution of phenotype is denoted μ′. Note that μ′ is greater than μ but less than μ_s. The quantity *S* is called the selection differential, and *R* is called the response to selection.

quantitative trait in a population, and the shaded part of the distribution to the right of the phenotypic value denoted T indicates those individuals selected for breeding. The value T is called the **truncation point**. The mean phenotype in the entire population is denoted μ, and that of the selected parents is denoted μ_s. When the selected parents are mated at random, their offspring have the phenotypic distribution shown in panel (b), where the mean phenotype is denoted μ'.

An example of truncation selection for seed weight in edible beans is shown in Figure 7. In this example, T = 650 mg, μ = 403.5 mg, μ_s = 691.7 mg, and μ' = 609.1 mg. In this case—as is typical of truncation selection—the offspring mean μ' is greater than the previous population mean μ but less than the parental mean μ_s. The reason μ' is greater than μ is that some of the selected parents have favorable genotypes and therefore pass favorable genes on to their offspring. At the same time, μ' is generally less than μ_s for two reasons:

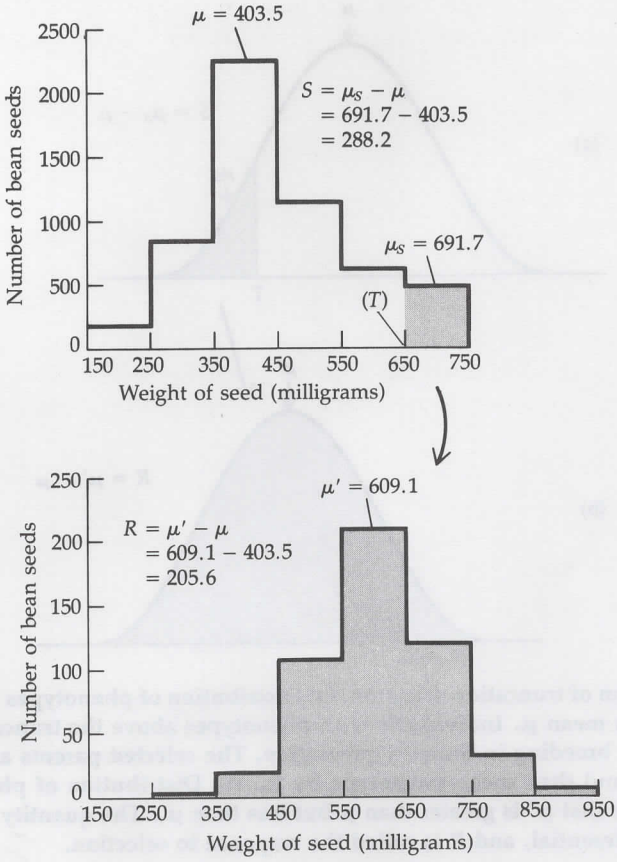


Figure 7. Truncation selection experiment for seed weight in edible beans of the genus *Phaseolus*, laid out as in Figure 6. The truncation point (T) is 650 mg. The selection differential S is the difference in means between the selected parents and the whole population. The response R is the difference in means between the progeny generation and the entire population in the previous generation. The quantity R/S is the realized heritability. (Data from Johannsen 1903.)

1. Because some of the selected parents do not have favorable genotypes; rather, their exceptional phenotypes result from chance exposure to exceptionally favorable environments.
2. Because alleles, not genotypes, are transmitted to the offspring, and exceptionally favorable genotypes are disrupted by Mendelian segregation and recombination.

The difference in mean phenotype between the selected parents and the entire parental population is the **selection differential** and is designated S. In symbols,

$$S = \mu_s - \mu \qquad \textbf{(8.8)}$$

The difference in mean phenotype between the progeny generation and the previous generation is the **response** to selection and is designated R. Symbolically,

$$R = \mu' - \mu \qquad (8.9)$$

In Figure 7, $\langle S \rangle = 288.2$ mg and $\langle R \rangle = 205.6$ mg.

In quantitative genetics, any equation that defines the relationship between the selection differential S and the response to selection R is known as a **prediction equation**. Since selection can be applied to a population in many different ways (others will be discussed later in this chapter), the prediction equation may differ corresponding to the different modes of selection. A general prediction equation that applies to many forms of selection, including truncation selection (the type of selection illustrated in Figure 6), is

$$R = h^2 S \qquad (8.10)$$

where h^2 is the **realized heritability**. Later in this chapter, we will show that the realized heritability is identical to the narrow-sense heritability defined by regression *provided* the phenotypes and the magnitudes of genetic effects follow a Gaussian distribution. These assumptions are necessary in order to apply regression to the problem. This equivalence emphasizes again that heritability can be understood at several different levels. Equation 8.10 implies that the realized heritability of a trait can be interpreted as a mere description of what happens when artificial selection is practiced. In Figure 7, for example, $\langle S \rangle = 288.2$ and $\langle R \rangle = 205.6$, so $\langle h^2 \rangle = \langle R \rangle / \langle S \rangle = 205.6/288.2 = 71.3\%$. When estimated like this from empirical data, h^2 is the realized heritability, and it simply summarizes the observed result.

Realized Heritability of Bristle Number in *Drosophila* D

Below are data on the number i of sternital bristles in samples from two consecutive generations G1 and G2 of an experiment in directional selection for increased bristle number. In the G1 generation, individuals with 22 or more bristles (enclosed in brackets) were mated together at random to form the G2 generation. Estimate the realized heritability of the number of sternital bristle in this experiment. (Data provided by T. Mackay. In order to make the sexes comparable, the value of two has been added to the bristle number in males.)

i	G1	G2	i	G1	G2	i	G1	G2
15	0	2	20	20	13	25	[1]	3
16	21	4	21	12	14	26	0	2
17	5	7	22	[13]	12	27	0	0
18	18	16	23	[3]	6	28	0	2
19	17	17	24	[5]	3			

ANSWER Estimates of the means are $\langle \mu \rangle = 2220/115 = 19.3$, $\langle \mu_S \rangle = 500/22 = 22.7$, $\langle \mu' \rangle = 2035/101 = 20.1$. Then $\langle S \rangle = 22.7 - 19.3 = 3.4$ (Equation 8.8), and $\langle R \rangle = 20.1 - 19.3 = 0.8$ (Equation 8.9). Realized heritability estimated from Equation 8.10 is $\langle h^2 \rangle = 0.8/3.4 = 23.5$ percent.

The data in Box D are from experiments by Mackay (1984) that demonstrate the potential significance of new mutations in quantitative genetics. The base population was formed by a cross which mobilizes the transposable element *P* and results in new insertions in the germ line and a syndrome of partial infertility and other reproductive abnormalities known as hybrid dysgenesis (discussed in Chapter 7). A genetically identical base population was formed by the reciprocal cross, in which the *P* element is not mobilized and hybrid dysgenesis does not occur. In the dysgenic cross, the realized heritability in abdominal bristle number was increased by 40% as compared with the non-dysgenic control. More strikingly, the phenotypic variance of bristle number in the selected dysgenic lines increased by a factor of three over the course of eight generations. These results demonstrate that the genetic variation affecting quantitative traits may even include insertions of transposable elements. On the other hand, some comparable experiments using hybrid dysgenesis have not given such dramatic results.

Selection Limits

Progress under artificial selection does not continue forever. Any population must eventually reach a selection limit, or **plateau**, after which it no longer responds to selection. One of the reasons that a population eventually reaches a plateau is exhaustion of genetic variance—when all alleles affecting the selected trait have become fixed, lost, or are otherwise unavailable for selection. With no genetic variance, no progress under individual selection can be achieved. However, many experimental populations that have reached a selection limit readily respond to reverse selection (selection in the reverse direction of that originally applied), so genetic variance affecting the trait is still present. Indeed, in such populations, the phenotype may change in the direction of its original value if continuing artificial selection is simply suspended (**relaxed selection**). The consequences of relaxed selection for a case in *Drosophila* are illustrated in Figure 8.

One frequent reason for the occurrence of selection limits in populations with considerable genetic variation is that artificial selection is opposed by natural selection. In mice, for example, response to selection for small body size ultimately ceases because small animals are less fertile than larger ones and the smallest animals are sterile (Falconer 1981). Selection for small body

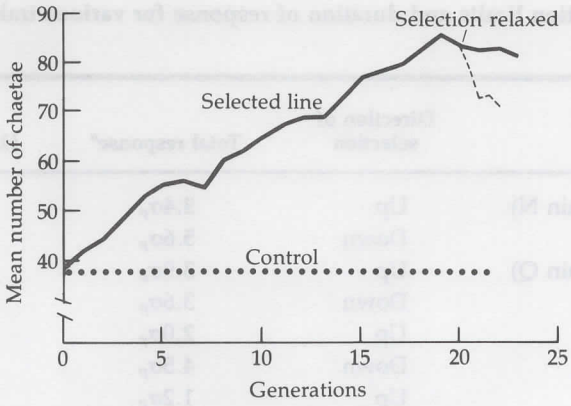


Figure 8. Response to selection for number of abdominal chaetae bristles in females of *Drosophila melanogaster* (solid line) and unselected control (no selection). Note the decrease in bristle number (dashed line) in the selected line upon relaxation of selection. (After Clayton et al. 1953.)

size gradually becomes less effective due to the opposing effects of natural selection until, eventually, no further progress is possible. When artificial selection is relaxed, the natural selection is unopposed and results in a retrogression in the artificially selected trait. Some backward slippage with relaxed selection also results from diminution in the linkage disequilibrium which usually builds up during the course of long-term artificial selection. If natural selection opposes the artificial selection, then when artificial selection is relaxed, natural selection results in at least a partial return to the initial phenotypic mean.

In most genetically heterogeneous populations, artificial selection can change the phenotype well beyond the range of variation found in the original population. Pertinent data for populations of mice are presented in Table 1. As can be seen, a total selection response of three to five times the original phenotypic standard deviation is not unusual, and for selection to change a population of effective size N halfway to its selection limit typically requires about $0.5N$ generations.

In some cases the total response to artificial selection is very large. For example, in a long-term selection experiment for pupal weight in *Tribolium*, in which the base population consisted of the progeny of a cross between two inbred lines, 100 generations of selection resulted in a population in which the mean pupal weight in the selected population was 17 standard deviation units greater than the mean in the base population (Enfield 1980).

Some traits consistently fail to respond to artificial selection, suggesting a lack of suitable genetic variation. Bilateral symmetry is an example of a trait

Table 1. Selection limits and duration of response for various traits in laboratory mice.

Character selected	Direction of selection	Total response[a]	Half-life of response[b]
Weight (in strain N)	Up	$3.4\sigma_p$	$0.6N$
	Down	$5.6\sigma_p$	$0.6N$
Weight (in strain Q)	Up	$3.9\sigma_p$	$0.2N$
	Down	$3.6\sigma_p$	$0.4N$
Growth rate	Up	$2.0\sigma_p$	$0.3N$
	Down	$4.5\sigma_p$	$0.5N$
Litter size	Up	$1.2\sigma_p$	$0.5N$
	Down	$0.5\sigma_p$	$0.5N$

(From Falconer 1977.)

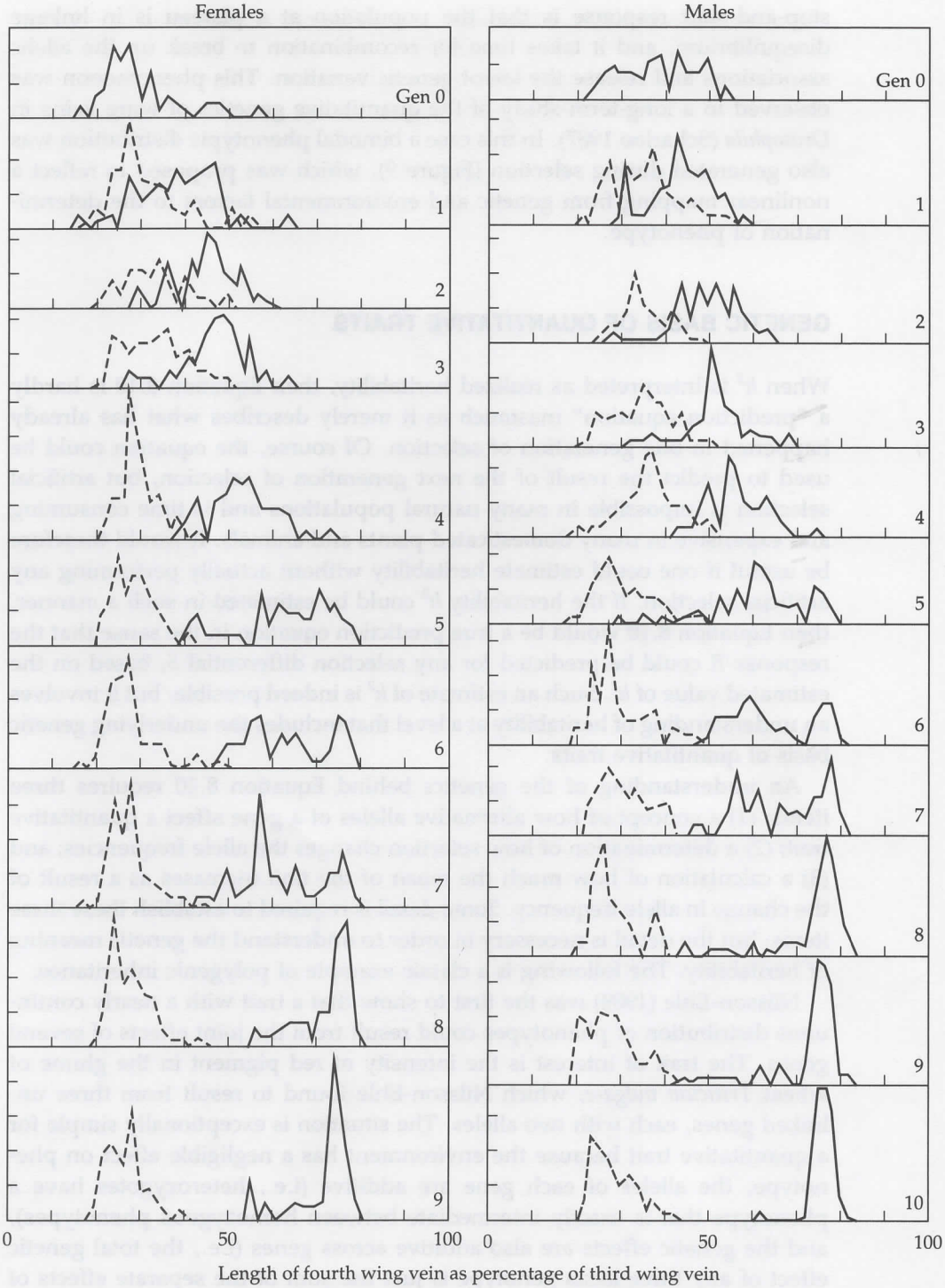
[a] Total response is expressed as a multiple of the initial phenotypic standard deviation, σ_p.

[b] Half-life of response is the number of generations taken to progress halfway to the selection limit; here the half-life is expressed in multiples of effective population number (N).

that has not been amenable to change by artificial selection. The failure of Maynard Smith and Sondhi (1961) to create bilateral asymmetry in *Drosophila* by selecting for an excess of dorsal bristles on the left side is typical. The apparent lack of genetic variation determining bilateral asymmetry is of interest in regard to development, for it implies that the genetic control of development of symmetrical structures specifies patterns that are common to the left and the right sides of the body. That is, rather than left-bristle genes and right-bristle genes, there appear to be generic bristle genes whose spatial expression is determined symmetrically.

Not all traits with heritable variation obey the prediction equation and show a simple linear change in the mean. Sometimes a trait responds to directional selection for a few generations, then ceases to respond, but later responds again as selection is continued. One possible mechanism for this

Figure 9. Frequency distributions in females (left) and males (right) of a line of *Drosophila melanogaster* selected for length of the fourth wing vein. The broken lines represent selection for a short vein, and solid lines represent selection for a long vein in separate populations. In the line selected for long veins, both sexes displayed a roughly bimodal frequency distribution when the relative vein length was approximately 60–80%. This result may suggest that there is a discontinuity either in the genetic basis for the trait, or in the developmental mapping from genotype to phenotype. (From Scharloo 1987.) ▷


Females
Males
Gen 0
1
2
3
4
5
6
7
8
9
10
0
50
100
Length of fourth wing vein as percentage of third wing vein

stop-and-start response is that the population at a plateau is in linkage disequilibrium, and it takes time for recombination to break up the allelic associations and release the latent genetic variation. This phenomenon was observed in a long-term study of the quantitative genetics of wing veins in *Drosophila* (Scharloo 1987). In this case a bimodal phenotypic distribution was also generated during selection (Figure 9), which was proposed to reflect a nonlinear mapping from genetic and environmental factors to the determination of phenotype.

GENETIC BASIS OF QUANTITATIVE TRAITS

When h^2 is interpreted as realized heritability, then Equation 8.10 is hardly a "prediction equation" inasmuch as it merely describes what has already happened in one generation of selection. Of course, the equation could be used to predict the result of the next generation of selection, but artificial selection is impossible in many natural populations and is time consuming and expensive in many domesticated plants and animals. It would therefore be useful if one could estimate heritability without actually performing any artificial selection. If the heritability h^2 could be estimated in such a manner, then Equation 8.10 would be a true prediction equation in the sense that the response R could be predicted for any selection differential S, based on the estimated value of h^2. Such an estimate of h^2 is indeed possible, but it involves an understanding of heritability at a level that includes the underlying genetic basis of quantitative traits.

An understanding of the genetics behind Equation 8.10 requires three items: (1) a concept of how alternative alleles of a gene affect a quantitative trait; (2) a determination of how selection changes the allele frequencies; and (3) a calculation of how much the mean of the trait increases as a result of the change in allele frequency. Some detail is required to establish these three items, but the detail is necessary in order to understand the genetic meaning of heritability. The following is a classic example of polygenic inheritance.

Nilsson-Ehle (1909) was the first to show that a trait with a nearly continuous distribution of phenotypes could result from the joint effects of several genes. The trait of interest is the intensity of red pigment in the glume of wheat *Triticum vulgare*, which Nilsson-Ehle found to result from three unlinked genes, each with two alleles. The situation is exceptionally simple for a quantitative trait because the environment has a negligible effect on phenotype, the alleles of each gene are additive (i.e., heterozygotes have a phenotype that is exactly intermediate between homozygous phenotypes), and the genetic effects are also additive across genes (i.e., the total genetic effect of any three-locus genotype is just the sum of the separate effects of

each gene). To simplify matters, consider just two of the genes, and let their alleles be denoted (*A*, *a*) and (*B*, *b*). With additivity within and across genes, we may assume that the genotype *aabb* has a color score of 0 (white) and that each *A* or *B* allele in the genotype contributes one unit of red pigment. Table 2 shows the nine possible two-gene genotypes, their frequencies with random mating when the allele frequencies of *A* and *B* are both 1/2, and the color score of each genotype assuming additivity. The mean color score of the population is 2. Indeed, when the allele frequencies of *A* and *B* are both *p*, then the mean of a population with random mating can be shown to equal $4p$. Suppose that the two lowest phenotypic classes (i.e., 0 and 1) are selected as parents of the next generation. Then we calculate μ_s, μ', S, R, and $h^2 = R/S$ by first calculating the allele frequency of *A* and *B* among selected parents, and then using the mean = $4p$ formula to obtain the mean of the offspring with random mating.

To be specific, $\mu = 4(1/2) = 2$ is given. The selected parents consist of genotypes *Aabb*, *aaBb*, and *aabb* with respective frequencies 2/5, 2/5, and 1/5, and the mean of parents = $\mu_s = (2/5)(1) + (2/5)(1) + (1/5)(0) = 4/5$. The allele frequency of *A* and *B* among parents = $(1/2)(2/5) = 1/5$, and therefore the mean among offspring is $\mu' = 4(1/5) = 4/5$. Then $S = (4/5) - 2 = -6/5$ and $R = (4/5) - 2 = -6/5$, so $h^2 = R/S = 1.0$. As demonstrated below, this high heritability is due to the additivity within and across genes and not merely to the fact that environmental effects are negligible.

The column headed "Phenotype with dominance" in Table 2 refers to a hypothetical situation in which the *A* and *B* alleles are dominant but still additive across genes. Thus, genotypes *AA*, *Aa*, *BB*, and *Bb* each add one unit of red pigment to the phenotype. In this case, it can be shown that the

Table 2. Additive and dominant phenotypes determined by two discrete loci.

Genotype	Frequency	Phenotype with additivity	Phenotype with dominance
AABB	1/16	4	2
AABb	2/16	3	2
AAbb	1/16	2	1
AaBB	2/16	3	2
AaBb	4/16	2	2
Aabb	2/16	1	1
aaBB	1/16	2	1
aaBb	2/16	1	1
aabb	1/16	0	0

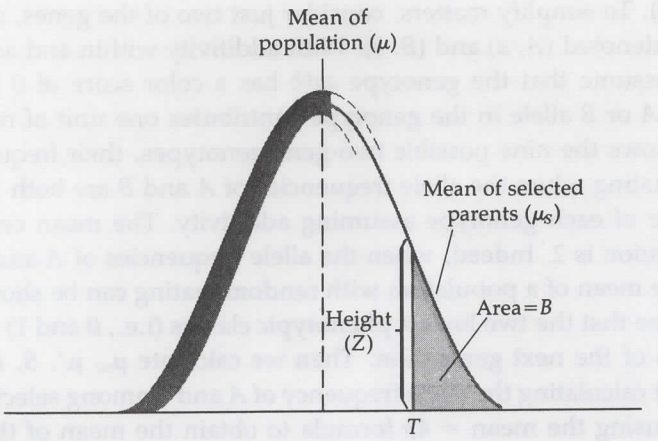


Figure 10. Normal distribution of a quantitative trait in a hypothetical population, showing some important symbols used in quantitative genetics. Here μ is the mean of the population, T the truncation point, Z the height (ordinate) of the normal density at the point T, B the shaded area under the normal curve to the right of T, and μ_s the mean among selected parents.

mean of a random-mating population with allele frequencies of A and B both equal to p is given by $2p(1 + q)$, where $q = 1 - p$. If the two lowest phenotypic classes (i.e., 0 and 1) are selected as parents of the next generation, then the mean of parents = μ_s = (1/7)(1) + (2/7)(1) + (1/7)(1) + (2/7)(1) + (1/7)(0) = 6/7. The allele frequency of A and B among parents is p = (1/7) + (1/2)(2/7) = 2/7, and the mean of the offspring is therefore μ' = 2(2/7)[1 + (5/7)] = 48/49. Thus, S = (6/7) − (3/2) = −9/14 and R = (48/49) − (3/2) = −51/98, so h^2 = 51/63 = 0.81.

Although environmental effects on seed color are still negligible in the dominance case, the heritability has become less than 1.0. This perhaps surprising result occurs because certain genetic effects (such as those resulting from dominance or, in other examples, nonadditivity across genes) are not useful in changing a population by means of the type of individual selection discussed here.

To see how an underlying genetic model can be formulated for continuous characters, refer to Figure 10, which shows the normal distribution of a trait in a hypothetical random mating population. In truncation selection, all individuals with phenotypes above the truncation point T are saved for breeding, and the shaded area B of the distribution represents the proportion of the population selected. (The total area under any normal density equals 1.) The height of the normal distribution at the point T is denoted Z, and, as before, the mean phenotype among the selected individuals is called μ_s.

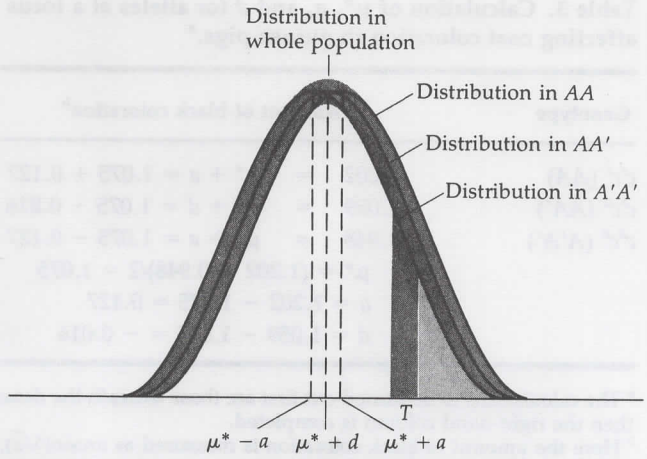


Figure 11. Same distribution as in Figure 10, showing the slightly different distributions of phenotypic value among the three genotypes (AA, AA', and $A'A'$) for the gene with two alleles that contributes to the quantitative trait. The means of the distributions of AA, AA', and $A'A'$ are symbolized $\mu^* + a$, $\mu^* + d$, and $\mu^* - a$, respectively.

One of the special properties of the normal distribution to be used below is that

$$(\mu_s - \mu)/\sigma^2 = Z/B \qquad (8.11)$$

To determine the amount of increase in mean phenotype in a population resulting from one generation of truncation selection, we first imagine a gene that affects the trait in question and that has alleles A and A' with respective allele frequencies p and q. Because of random mating, genotypes AA, AA', and $A'A'$ are present in the population with frequencies p^2, $2pq$, and q^2, respectively, but the individual genotypes cannot be identified through their phenotypic values because of the variation in phenotype caused by environmental factors and genetic differences in other genes. If the genotypes could be identified, their individual distributions of phenotypic value might appear as shown in Figure 11. Each distribution is normal and has the same variance, but the means are slightly different. The mean phenotypes of AA, AA', and $A'A'$ genotypes are denoted $\mu^* + a$, $\mu^* + d$, and $\mu^* - a$, respectively. The symbols a and d serve as convenient representations of the effects of the alleles in question on the quantitative trait. The difference between means of homozygotes is $(\mu^* + a) - (\mu^* - a) = 2a$, and d/a serves as a measure of dominance. The relationship $d = a$ means that A is dominant, $d = 0$ implies additivity (i.e., heterozygotes exactly intermediate in phenotype between the homozygotes), and $d = -a$ means that A' is dominant. (Use of a and d in

Table 3. Calculation of μ^*, a, and d for alleles at a locus affecting coat coloration in guinea pigs.[a]

Genotype	Amount of black coloration[b]
$c^r c^r$ (AA)	$1.202 = \mu^* + a = 1.075 + 0.127$
$c^r c^d$ (AA')	$1.059 = \mu^* + d = 1.075 - 0.016$
$c^d c^d$ ($A'A'$)	$0.948 = \mu^* - a = 1.075 - 0.127$
	$\mu^* = (1.202 + 0.948)/2 = 1.075$
	$a = 1.202 - 1.075 = 0.127$
	$d = 1.059 - 1.075 = -0.016$

[a] The calculations to be carried out first are those beneath the data; then the right-hand column is completed.
[b] Here the amount of black coloration is measured as $\arcsin(\sqrt{x})$, where x is the percentage of black coloration on the animal. For $c^r c^r$, $c^r c^d$, and $c^d c^d$ genotypes, the corresponding x values are 87%, 76%, and 66%, respectively.
(Data from Wright 1968.)

this manner simplifies some of the subsequent formulas.) Calculation of a and d for an actual example involving two alleles that affect coat coloration in guinea pigs is illustrated in Table 3. In this case, $\langle a \rangle = 0.127$, $\langle d \rangle = -0.016$ (the negative sign on d means that the c^d allele is partially dominant), and $\langle d/a \rangle = -0.126$. Referring back to Figure 11, the mean phenotype in the entire population is

$$\mu = p^2(\mu^* + a) + 2pq(\mu^* + d) + q^2(\mu^* - a) \qquad \textbf{(8.12)}$$

Effects of the scale of measurement are known as *scaling effects*. For example, the a and d values in Table 3 are different when calculated for the percent of black coloration x or for the values of $\arcsin(\sqrt{x})$ tabulated. Since estimates of the additive and dominance values of alleles depend on scaling, so does the heritability. An important point is that the equivalence between the heritability defined by parent-offspring regression and by realized heritability depends on the correct choice of scaling. Only one scaling provides a Gaussian distribution of phenotypes and of genetic effects, and that is the scaling that yields the prediction equation.

In the next section we calculate the change in allele frequency that occurs as a result of directional selection.

Change in Gene Frequency

Suppose for the moment that we were practicing artificial selection for increased amount of black coat coloration in the guinea pigs in Table 3. Selection

for black coat coloration in a population containing both the c^r (i.e., A) and c^d (i.e., A') alleles would be successful in increasing the allele frequency of A, and the average amount of black coloration among individuals of the next generation would increase. Therefore, in order to calculate the expected increase in black coloration in one generation of selection, we must first calculate the corresponding change in the allele frequency of A. An equation for change in allele frequency with natural selection was derived in Chapter 4, which remains valid for artificial selection if we agree to interpret the "fitness" of an individual as the probability that the individual is included among the group selected as parents of the next generation. With this interpretation of fitness, differences in fitness (i.e., reproductive success) of AA, AA', and $A'A'$ genotypes correspond to the differences in area to the right of the truncation point in Figure 11, because only those individuals in the shaded area are allowed to reproduce. The differences in area are easy to calculate if you shift or slide each curve horizontally until its mean coincides with μ^*. The $A'A'$ curve must slide a units to the right, and the AA' and AA curves must slide d and a units to the left. This shifting brings the distributions into coincidence, but it slides the truncation points slightly out of register, as shown in Figure 12. The difference in "fitness" between AA and AA', denoted $w_{11} - w_{12}$ (as in Chapter 4), is equal to the small area indicated

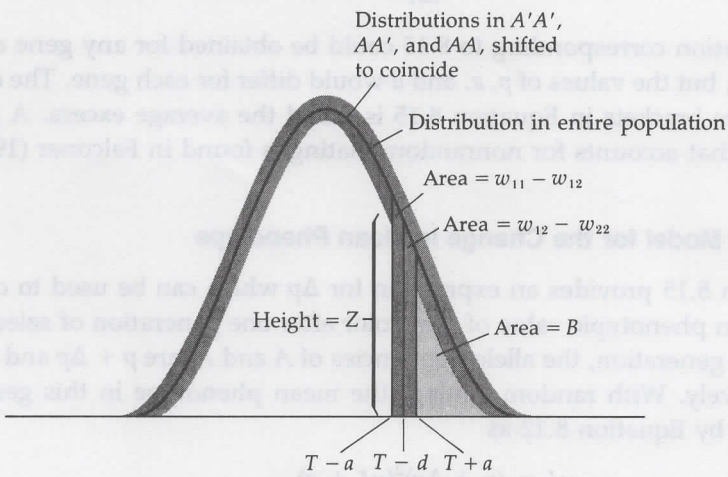


Figure 12. Same distribution as in Figures 10 and 11, but with the distributions of AA, AA', and $A'A'$ shifted laterally to coincide. Shifting the distributions slides the truncation points slightly out of register, so the truncation points for AA, AA', and $A'A'$ become $T - a$, $T - d$, and $T + a$, respectively. The small area that is denoted $w_{11} - w_{12}$ is the difference between the proportions of AA and AA' genotypes that are included among the selected parents, and the area $w_{12} - w_{22}$ is the difference in the proportion of AA' and $A'A'$ genotypes included among the selected parents.

in Figure 12; as is the difference in fitness between AA' and $A'A'$, denoted $w_{12} - w_{22}$. The areas corresponding to $w_{11} - w_{12}$ and $w_{12} - w_{22}$ are approximately rectangles, and the area of a rectangle is the product of the base and the height. The approximation is most accurate when the effect of this one locus on the phenotype is small. Therefore, since Z represents the height of the normal distribution at the point T, we can make the following approximations:

$$\begin{aligned} w_{11} - w_{12} &\approx Z[(T - d) - (T - a)] = Z(a - d) \\ w_{12} - w_{22} &\approx Z[(T + a) - (T - d)] = Z(a + d) \end{aligned} \tag{8.13}$$

The average fitness $\overline{w}$ of the entire population simply equals B, because B is the proportion of the population saved for breeding. From Equation 4.4 we know that

$$\Delta p = pq[p(w_{11} - w_{12}) + q(w_{12} - w_{22})]/\overline{w},$$

where Δp is the change in frequency of the allele A in one generation of selection. Substituting from Equation 8.13 and using $\overline{w} = B$ leads to

$$\Delta p = pq[pZ(a - d) + qZ(a + d)]/B \tag{8.14}$$

or, since $p + q = 1$,

$$\Delta p = \left(\frac{Z}{B}\right)pq[a + (q - p)d] \tag{8.15}$$

An equation corresponding to 8.15 could be obtained for any gene affecting the trait, but the values of p, a, and d would differ for each gene. The quantity in square brackets in Equation 8.15 is called the average excess. A generalization that accounts for nonrandom mating is found in Falconer (1985).

Genetic Model for the Change in Mean Phenotype

Equation 8.15 provides an expression for Δp which can be used to calculate the mean phenotypic value of coat color after one generation of selection. In the next generation, the allele frequencies of A and A' are $p + \Delta p$ and $q - \Delta p$, respectively. With random mating, the mean phenotype in this generation is given by Equation 8.12 as

$$\begin{aligned} \mu' = {} & (p + \Delta p)^2(\mu^* + a) \\ & + 2(p + \Delta p)(q - \Delta p)(\mu^* + d) \\ & + (q - \Delta p)^2(\mu^* - a) \end{aligned} \tag{8.16}$$

When the right side of this expression is multiplied out, ignoring $(\Delta p)^2$ because Δp is usually very small, μ' is found to approximate

$$\mu' \simeq \mu + 2[a + (q - p)d]\Delta p \tag{8.17}$$

The approximation in Equation 8.17 is rather good even for relatively large values of Δp.

Equation 8.17 warrants a little more development since it yields the prediction equation $R = h^2S$ (Equation 8.10) and also provides an expression for h^2 in terms of the parameters a, d, and p that can be interpreted genetically. First, rewrite Equation 8.17 as

$$\mu' - \mu = 2[a + (q - p)d]\Delta p \tag{8.18}$$

Then substitute for Δp from Equation 8.15, which yields

$$\mu' - \mu = (Z/B)2pq[a + (q - p)d]^2 \tag{8.19}$$

Now use the expression for Z/B given in Equation 8.11 to obtain

$$\mu' - \mu = (\mu_s - \mu)2pq[a + (q - p)d]^2/\sigma^2 \tag{8.20}$$

Finally, substitute from Equations 8.8 and 8.9 for the selection differential S and the response R, which yields

$$R = (S)2pq[a + (q - p)d]^2/\sigma^2 \tag{8.21}$$

However, $R = h^2S$ also (Equation 8.10), so

$$h^2 = 2pq[a + (q - p)d]^2/\sigma^2 \tag{8.22}$$

Equation 8.22 for h^2 is the one we were after, as it defines the heritability in terms of p, q, a, and d—each of which has a genetic meaning.

Equation 8.22 is a valid approximation when a single gene affects the trait in question, and when the effects of that gene are small. However, when many genes affect the trait, the right side of the equation must be replaced by a summation of such terms, one for each gene. That is, for many genes, $R = h^2S$, where

$$h^2 = \frac{\Sigma 2pq[a + (q - p)d]^2}{\sigma^2} \tag{8.23}$$

in which the summation is over all genes that affect the trait. (However, each gene may have different values of a, d, p, and q.) As will be discussed in more detail later, the quantity

$$\sigma_a^2 = \Sigma 2pq[a + (q - p)d]^2 \tag{8.24}$$

is called the **additive genetic variance** of the trait. Although the individual components in the additive genetic variance are difficult to identify except in contrived examples like the one involving guinea pigs, the collective effects (represented by the summation) can be estimated.

COMPONENTS OF PHENOTYPIC VARIANCE

As Equation 8.24 implies, the variance of a quantitative trait can be split into various components representing different causes of variation. Similarity between relatives is conveniently expressed in terms of the variance components, but variance partitioning is also of interest in its own right. Since the rate of change of a trait under selection depends on the amount of genetic variation affecting the trait, if there is no genetic variation, there is obviously no response to selection. What is not so obvious is that some components of genetic variation cannot be acted upon by some kinds of selection. In other words, certain populations have ample genetic variation, yet fail to respond to selection. The part of the genetic variation amenable to selection is clarified by partitioning the variance.

Genetic and Environmental Sources of Variation

Table 4 shows the kind of reasoning involved in partitioning the variance. The phenotypic value of any individual is represented as a sum of three components: (1) the mean μ of the entire population, (2) a deviation from the population mean due to the specific genotype of the individual in question (symbolized as G_1, G_2, and G_3 for AA, AA', and $A'A'$ genotypes, respectively), and (3) a deviation from the population mean due to the specific

Table 4. Phenotypes of various genotypes as the sum of μ, G, and E.[a]

Genotype	Phenotypic value
AA	$\mu + G_1 + E_1$
AA	$\mu + G_1 + E_2$
AA	$\mu + G_1 + E_3$
AA'	$\mu + G_2 + E_4$
AA'	$\mu + G_2 + E_5$
AA'	$\mu + G_2 + E_6$
$A'A'$	$\mu + G_3 + E_7$
$A'A'$	$\mu + G_3 + E_8$
$A'A'$	$\mu + G_3 + E_9$

[a] μ is the population mean. G is a contribution due to genotype, different for each genotype. E is a contribution due to environment, different for each individual.

microenvironment of the individual in question (the environmental deviations are unique to each individual and are represented as $E_1, E_2, \ldots, E_9$). These microenvironmental effects might be due to random differences in nutrition, temperature, or other external factors, or they might be seen even in an absolutely uniform external environment due to the vagaries of development. It is important to note that the G's and E's are not directly observable. Nevertheless, as we shall see, the total variance in phenotypic value can be partitioned into a component due to variation among the G's and another component due to variation among the E's. The model in Table 4 can be summarized by writing

$$P = \mu + G + E \qquad \textbf{(8.25)}$$

where P represents the phenotypic value of any individual and G and E are the genotypic and environmental deviations pertaining to that individual.

To connect the above symbols with actual numbers, we may use Table 3, and assume an allele frequency of $p = 0.2$ in a hypothetical population. The mean of the population is determined from Equation 8.12 as $\mu = 0.994$. Thus, the respective G_1, G_2, and G_3 deviations for AA, AA', and $A'A'$ genotypes are

$$G_1 = 1.202 - 0.994 = 0.208$$
$$G_2 = 1.059 - 0.994 = 0.065$$
$$G_3 = 0.948 - 0.994 = -0.046$$

For a particular animal of genotype AA whose actual coat color score is, for example, 1.312, the corresponding value of E for the animal would be calculated using Equation 8.25 from the expression $1.312 = 0.994 + 0.208 + E$; thus, for this animal, $E = 0.110$. Similarly, a particular animal of genotype AA' with an actual phenotype of $P = 1.009$ would have a value of E given by $1.009 = 0.994 + 0.065 + E$, or $E = -0.050$. Because the E values are defined as deviations from their mean, the average of E's for any genotype is 0. Likewise, since the G's are defined as deviations from their mean, the mean of the G's is 0, which can be verified in the guinea pig example because

$$(0.2)^2G_1 + 2(0.2)(0.8)G_2 + (0.8)^2G_3 = 0$$

Equation 8.25 is appropriate when the effects of genotype and environment are additive—that is, when the deviation of the phenotype of any particular individual from the population mean ($P - \mu$) can be written as the sum of an effect resulting from the genotype of that individual and a separate effect resulting from the environment of that individual.

To this point the discussion has been restricted to a particular population in a single **macroenvironment**, and the sources of variation have been due to genetic and microenvironmental differences among individuals. A change

in macroenvironment is easiest to see in an experimental setting, where for example, in macroenvironment 1 all of the guinea pigs get twice as much food as in macroenvironment 2. Microenvironmental variation occurs within each macroenvironment, and might include such things as on which side of the cage a guinea pig was born, or how much nesting material it finds. Additivity of genetic and environmental effects is true whenever the ratio of G_1:G_2:G_3 is the same in each of the relevant environments. For the genotypes in Figure 13, for example, if the actual range of environments is the range designated E_1, then the genetic and environmental effects are additive because the ratio G_1:G_2:G_3 is the same for any particular environment in E_1. For the same reason, the genetic and environmental effects are additive if the actual range of environments is E_2. However, if the actual range of environments includes both E_1 and E_2, then the ratio G_1:G_2:G_3 depends on the particular environment, and therefore the genetic and environmental effects may not be additive. Nonadditivity of genetic and environmental effects is called **genotype–environment interaction**, and in writing Equation 8.25, it appears that there is an assumption that there is no genotype-environment interaction. In actually estimating components of variance, it is not necessary to assume that there is no genotype-environment interaction, because we can explicitly examine the phenotypes when reared in different macroenvi-

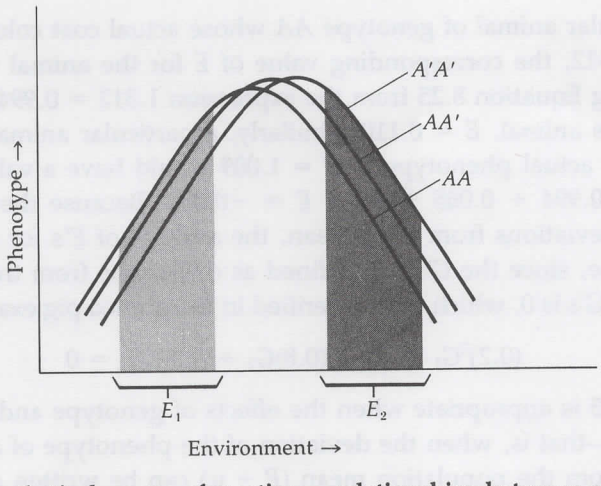


Figure 13. Hypothetical norms of reaction — relationships between phenotype and environment — for genotypes *AA*, *AA'*, and *A'A'*. In the range of environments denoted E_1, *A* is very nearly dominant to *A'* (that is, *AA* and *AA'* have nearly the same phenotype). However, in the range E_2, *A* and *A'* are very nearly additive (no dominance). The heritability of the trait resulting from this gene differs according to whether the population is reared in E_1 or E_2 environments.

ronments and directly estimate the magnitude of the interaction. Alternatively we can simply define the environmental variance to include the effects of genotype-environment interaction.

When Equation 8.25 is valid, the total phenotypic variance σ_P^2 in the population equals the mean of $(P - \mu)^2$. However, Equation 8.25 says that $(P - \mu)^2$ equals $(\mu + G + E - \mu)^2$, which is

$$\sigma_P^2 = (G + E)^2 = G^2 + 2GE + E^2 \qquad \textbf{(8.26)}$$

Because G and E are already deviations from their means, the mean of G^2 is the phenotypic variance in the population resulting from differences in genotype, and the mean of E^2 is the phenotypic variance resulting from differences in environment. The mean of G^2 is called the **genotypic variance** and is denoted σ_g^2. The mean of E^2 is called the **environmental variance** and is denoted σ_e^2. The remaining term—the mean of $2GE$—is two times the **genotype-environment covariance**. If the genotypic and environmental deviations are uncorrelated—that is, if there is no systematic association between genotype and environment—then there is said to be no genotype-environment association and the mean of $2GE$ equals zero. When there is no genotype-environment association, therefore,

$$\sigma_P^2 = \sigma_g^2 + \sigma_e^2 \qquad \textbf{(8.27)}$$

Equation 8.27 is the theoretical foundation for partitioning the variance into genetic and environmental effects. The assumption that genotype–environment association is negligible is frequently a valid assumption in animal and plant breeding where, because breeders have a degree of control not available to, for example, human geneticists, experiments can be intentionally designed in such a way as to minimize genotype-environment association. However, genotype-environment association can occur even in animal and plant breeding. For example, dairy farmers routinely provide more feed supplements to cows that produce more milk; because the ability to produce milk is partly due to genotype, this feed regimen will provide superior environments (i.e., better feed) to cows that have superior genotypes to begin with, so there will be a genotype–environment association. Similarly the best race horses get the best trainers and the children of the best students often go to the best schools. If one is not careful, genotype–environment association can inflate the apparent σ_g^2 and possibly give spurious overestimates of heritability.

The biological meaning of Equation 8.27 is shown for the alleles of one gene in Figure 14. The solid curves represent the phenotypic distributions in the genotypes AA, AA', and $A'A'$ with means denoted G_1, G_2, and G_3, and the dashed curve represents the phenotypic distribution in the entire population. The total phenotypic variance σ_P^2 is the variance of the dashed distri-

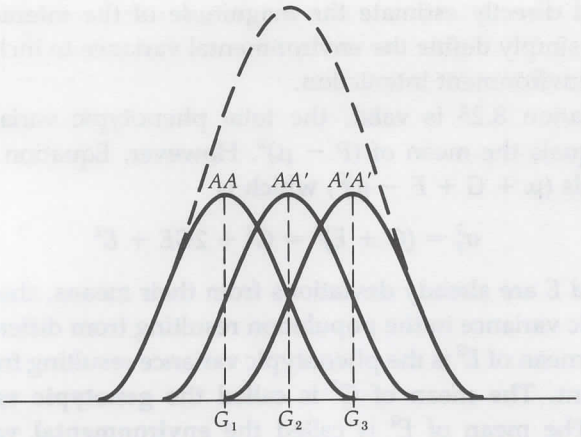


Figure 14. Phenotypic distribution (dashed curve) of a quantitative trait in a hypothetical population, showing distributions (solid curves) of three constituent genotypes for two alleles of a gene. The means of *AA*, *AA'*, and *A'A'* genotypes are denoted G_1, G_2, and G_3, respectively.

bution; the genotypic variance σ_g^2 is the variance among the G's (i.e., $\sigma_g^2 = p^2G_1^2 + 2pqG_2^2 + q^2G_3^2$, where p is the allele frequency of A); and the environmental variance σ_e^2 is obtained by subtraction: $\sigma_e^2 = \sigma_p^2 - \sigma_g^2$. Although the Gs are not generally known, σ_g^2 must equal zero in a genetically uniform population. The observed variance of a randomly bred population, therefore, provides an estimate of $\sigma_g^2 + \sigma_e^2$, whereas the average observed variance of genetically uniform populations provides an estimate of σ_e^2. The estimate of σ_g^2 can be obtained by subtraction as noted in an example using thorax length in *Drosophila* (shown in Table 5). In this case, genetic variation among individuals in the randomly bred population accounts for about 0.180/0.366 = 49.2% of the phenotypic variance. Genetically uniform populations such as inbred lines or crosses between inbreds are not available in human populations, of course, but identical twins are often used instead because of the identical genotypes of the twins.

An example of a naturally occurring organism that exhibits remarkably low levels of genetic variability is the African cheetah (O'Brien et al. 1983). One might suppose that limited genetic variation would result in depressed phenotypic variability as well, but a study of cranial measures by Wayne et al. (1986) revealed that the amount of variability was not appreciably less than that in three other large cats. In fact, there was a significant *increase* in the amount of fluctuating asymmetry (that is, the difference in measurements from the left and the right side of the body). The fluctuating asymmetry is consistent with the notion that genetic homozygosity results in reduced developmental stability—an idea that has considerable empirical support, but

Table 5. Calculation of genotypic variance (σ_g^2) and environmental variance (σ_e^2).[a]

Variance	POPULATIONS	
	Random-bred	Uniform
Theoretical	$\sigma_g^2 + \sigma_e^2$	σ_e^2
Observed	0.366	0.186

$\sigma_e^2 = 0.186$
$\sigma_g^2 = (\sigma_g^2 + \sigma_e^2) - \sigma_e^2 = 0.366 - 0.186 = 0.180$

[a] Trait is length of thorax in *Drosophila melanogaster* (in units of 10^{-2} mm). Data from Robertson (1957).

so far no good explanation in molecular terms. In any event, reduction of genetic variance, and concomitant high homozygosity, can result in an increase in phenotypic variance due to developmental instability. Since extreme homozygosity may result in phenotypes that are very sensitive to environmental fluctuations, the paradoxical increase in phenotypic variance results from genotype-environment interaction.

Components of Genotypic Variation

So far, the phenotypic variance has been partitioned into the genotypic variance and the environmental variance, according to Equation 8.27. The genotypic variance can be partitioned further into terms that are particularly important for interpreting the resemblance between relatives. The appropriate model is shown in Table 6, where the phenotypic means of AA, AA', and $A'A'$ are denoted in column 3 by $\mu^* + a$, $\mu^* + d$, and $\mu^* - a$, as they were earlier in Figure 11. To obtain the G values, the mean of each genotype must be expressed as a deviation from the population mean, which is $\mu = \mu^* + (p - q)a + 2pqd$, and the deviations are shown in the last column of Table 6. The genotypic variance σ_g^2 works out to be

$$\begin{aligned}\sigma_g^2 &= p^2G_1^2 + 2pqG_2^2 + q^2G_3^2 \\ &= 2pq[a + (q - p)d]^2 + (2pqd)^2\end{aligned} \tag{8.28}$$

The first term in Equation 8.28 is the additive genetic variance σ_a^2 encountered earlier in Equation 8.24. The second term is a new quantity called the **dominance variance**, which is symbolized σ_d^2. From Equation 8.28, therefore,

$$\sigma_g^2 = \sigma_a^2 + \sigma_d^2 \tag{8.29}$$

Table 6. Expressions for population mean and genotypic deviations.

Genotype	Frequency	Mean phenotype	Genotypic deviation from population mean (G)
AA	p^2	$\mu^* + a$	$G_1 = \mu^* + a - \mu = 2q[a + (q - p)d] - 2q^2d$
AA'	$2pq$	$\mu^* + d$	$G_2 = \mu^* + d - \mu = (q - p)[a + (q - p)d] + 2pqd$
$A'A'$	q^2	$\mu^* - a$	$G_3 = \mu^* - a - \mu = -2p[a + (q - p)d] - 2p^2d$

$$\begin{aligned}\text{Population mean } (\mu) &= p^2(\mu^* + a) + 2pq(\mu^* + d) + q^2(\mu^* - a)\\ &= (p^2 + 2pq + q^2)\mu^* + (p^2 - q^2)a + 2pqd\\ &= (p + q)^2\mu^* + (p - q)(p + q)a + 2pqd\\ &= \mu^* + (p - q)a + 2pqd\end{aligned}$$

which allows us to express the total phenotypic variance as the sum of three terms, namely

$$\sigma_p^2 = \sigma_a^2 + \sigma_d^2 + \sigma_e^2 \qquad \textbf{(8.30)}$$

When Equation 8.22 for heritability is written in terms of variance components rather than p, q, a and d, the equation implies that

$$h^2 = \sigma_a^2/\sigma_p^2 \qquad \textbf{(8.31)}$$

Equation 8.31 is an important result because it states that the heritability depends only on the additive genetic variance and not on the dominance variance. Therefore, if all the genetic variance in a population results from dominance variance (i.e., $\sigma_a^2 = 0$), then the population cannot respond to individual selection because h^2 equals zero. To say the same thing in another way, the dominance variance σ_d^2 represents that portion of the genetic variance that is not acted upon by individual selection.

Equation 8.31 means that the heritability of a trait is the ratio of the additive genetic variance to the total phenotypic variance. Sometimes the word *heritability* is used in reference to a different variance ratio, namely the ratio of the total genotypic variance to the total phenotypic variance (i.e., σ_g^2/σ_p^2). To avoid confusion, quantitative geneticists distinguish the two types of heritability as follows:

1. The ratio σ_a^2/σ_p^2 is called heritability in the *narrow sense*
   (This is the variance ratio we have been using all along.)
2. The ratio σ_g^2/σ_p^2 is called heritability in the *broad sense*.

Generally speaking, narrow-sense heritability is the more important with individual selection (or any mode of selection that capitalizes primarily on

the additive genetic variance), whereas broad-sense heritability is the more important when selection is practiced among clones (a clone is a group of genetically identical individuals), inbred lines, or varieties. We use the term *heritability* to mean narrow-sense heritability unless otherwise stated.

Heritability of Simple Mendelian Traits

E

By definition, a simple Mendelian trait is one that is determined entirely by genotype in the prevailing environment. Therefore, $\sigma_e^2 = 0$ in Equation 8.27, and the broad-sense heritability $\sigma_g^2/\sigma_p^2 = 1$. Show that, for a simple Mendelian recessive, the narrow-sense heritability equals $2q/(1 + q)$, where q is the recessive allele frequency.

ANSWER Let the phenotypes of AA, AA', and $A'A'$ be assigned phenotypic values 0, 0, and 1, respectively, so that the A' allele is recessive. Therefore, $\mu^* = 1/2$, $a = -1/2$, and $d = -1/2$. The numerator of Equation 8.31 is the additive genetic variance in phenotypic value, which equals $2pq^3$. Then the mean phenotype equals q^2, and the variance in phenotypic value is $q^2 - (q^2)^2 = q^2(1 - q^2) = q^2(1 - q)(1+q) = pq^2(1+q)$. Therefore, the heritability of the autosomal recessive trait is $h^2 = 2q/(1+q)$. When the autosomal recessive trait is rare, $q \simeq 0$, and the heritability is approximately equal to the frequency of heterozygous carriers.

As emphasized earlier, heritability has no transparent interpretation in simple genetic terms. The same is true of the variance components σ_a^2 and σ_d^2. Even for a single gene, the variance components depend on the particular values of a and d (Figure 15), and of course the estimates of heritability must also depend on allele frequency (Figure 16). With many genes which act together, σ_a^2 is defined as a summation of the values of $2pq[a + (q - p)d]^2$ for each gene affecting the trait, and σ_d^2 represents a summation of the values of $(2pqd)^2$ for each gene. Indeed, when the trait is affected by multiple genes, the formula for σ_g^2 in Equation 8.29 must be extended to include an additional term that pertains to interaction among the genes. This interaction term is called the **interaction variance** or the **epistatic variance** and is symbolized σ_i^2. With the interaction variance included, Equation 8.29 becomes

$$\sigma_g^2 = \sigma_a^2 + \sigma_d^2 + \sigma_i^2 \qquad \textbf{(8.32)}$$

The important point to remember about the components of genotypic variance is that they represent the cumulative, statistical effects of all genes affecting the trait. Few inferences about the actual mode of inheritance of the trait are possible from the variance components, particularly concerning the number of genes involved and their individual effects.

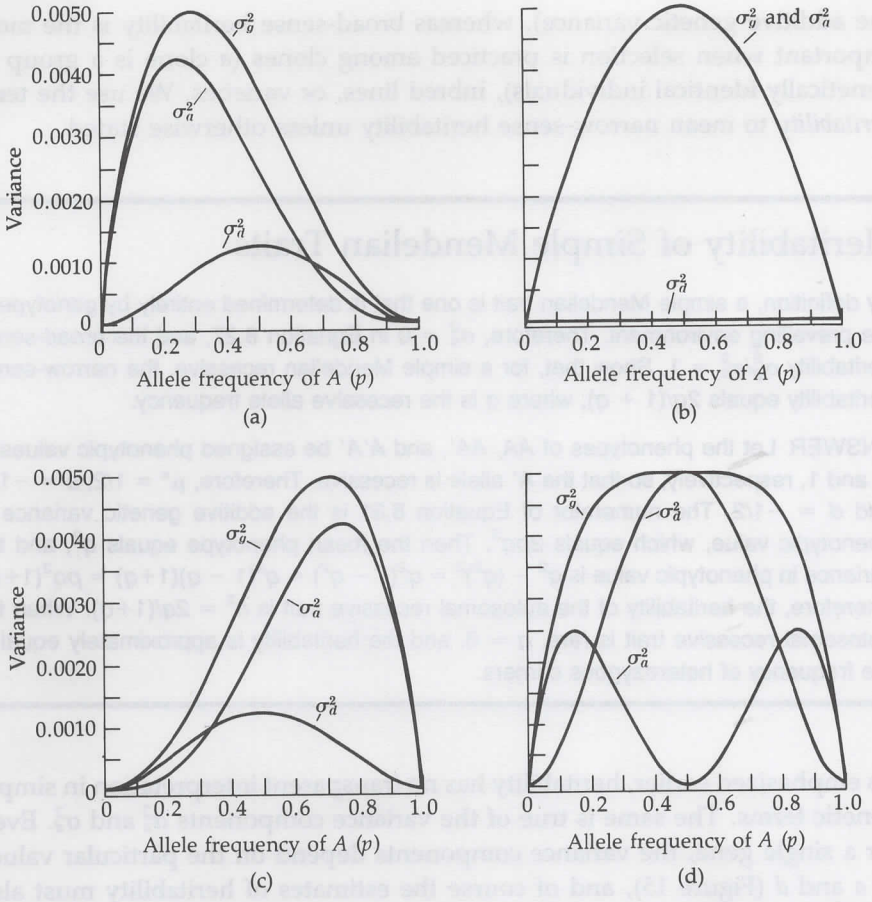


Figure 15. Total genetic variance (σ_g^2), additive genetic variance (σ_a^2), and dominance variance (σ_d^2) for a locus with two alleles (A and A'), plotted against the allele frequency of A (p). The mean phenotypes of AA, AA', and $A'A'$ are denoted $\mu^* + a$, $\mu^* + d$, and $\mu^* - a$, respectively. In all cases $\sigma_a^2 = 2pq[a + (q - p)d]^2$, $\sigma_d^2 = (2pqd)^2$, and $\sigma_g^2 = \sigma_a^2 + \sigma_d^2$. (a) $a = d = 0.0701$ (A dominant to A'); (b) $a = 0.1$, $d = 0$ (no dominance); (c) $-a = d = 0.07071$ (A' dominant to A); (d) $a = 0$, $d = 0.14142$ (overdominance). For ease of comparison, the values of a and d have been chosen to make the maximum of σ_g^2 equal to 0.005 in each case.

COVARIANCE BETWEEN RELATIVES

Components of genetic variation are important because they may be used to express the phenotypic covariance between relatives. In order to establish these relationships, we will assume that the population is in Hardy-Weinberg proportions. Since the distribution of offspring from a given parental genotype depends on the distribution of potential mates, the variance components

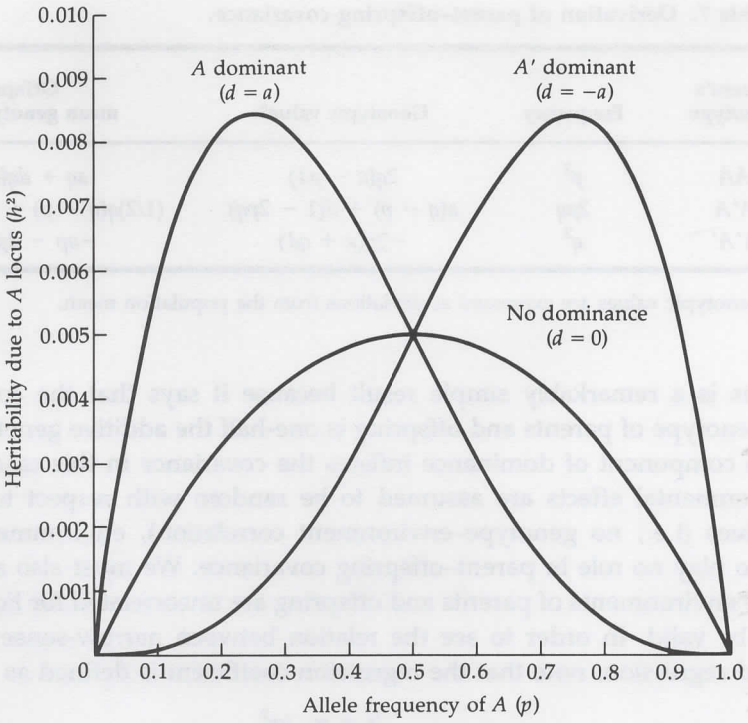


Figure 16. Narrow-sense heritability due to a single locus with two alleles (A and A') as a function of p, the allele frequency of A. In general, for one locus, $h^2 = 2pq[a + (q - p)d]^2/\sigma^2$, where σ^2 is the total phenotypic variance. The curves correspond to $a = 0.1$ and $d = 0.1$ (A dominant), $d = 0$ (no dominance), and $d = -0.1$ (A' dominant).

and estimates of heritability depend not only on allele frequencies but also on the distribution of genotype frequencies. To simplify things, we assume that the trait is determined by one gene with two alleles, and that the population is in Hardy-Weinberg proportions. However, the same results are also true for many genes when the trait is determined by summing the individual allelic effects, and the population is in multilocus linkage equilibrium.

Table 7 displays three genotypes of parents, their genotypic value, and the mean genotypic value of the offspring with random mating. The covariance of the offspring and one parent is calculated by taking the product of the last three columns of Table 7 and subtracting the product of the means of the last two columns. After tedious algebra, the covariance of offspring and parent is:

$$\sigma_{OP} = pq\,[a + (q - p)d]^2 = (1/2)\,\sigma_a^2 \qquad \textbf{(8.33)}$$

Table 7. Derivation of parent–offspring covariance.

Parent's genotype	Frequency	Genotypic value[a]	Offspring mean genotypic value
AA	p^2	$2q(a - pd)$	$aq + dq(q - p)$
$A'A$	$2pq$	$a(q - p) + d(1 - 2pq)$	$(1/2)q(q - p) + (1/2)d(q - p)^2$
$A'A'$	q^2	$-2p(a + qd)$	$-ap - dp(q - p)$

[a] Genotypic values are expressed as deviations from the population mean.

This is a remarkably simple result because it says that the covariance in phenotype of parents and offspring is one-half the additive genetic variance. No component of dominance inflates the covariance in this case. Since environmental effects are assumed to be random with respect to genotypic values (i.e., no genotype–environment correlation), environmental effects also play no role in parent–offspring covariance. We must also assume that the environments of parents and offspring are uncorrelated for Equation 8.33 to be valid. In order to see the relation between narrow-sense heritability and regression, note that the regression coefficient is defined as

$$b = \sigma_{xy}/\sigma_x^2 \tag{8.34}$$

In this case, the regression of offspring on one parent is, from Equation 8.33:

$$\begin{aligned} b_{OP} &= \sigma_{OP}/\sigma_p^2 \\ &= (1/2)\sigma_a^2/\sigma_p^2 \\ &= (1/2)h^2 \end{aligned} \tag{8.35}$$

In the regression of offspring on the *midparent* (that is, the average value of the parents), the denominator in Equation 8.35 becomes $(1/2)\sigma_p^2$ because it is the variance of the mean of the two parents, assuming random mating. Hence the regression coefficient equals $(1/2)\sigma_a^2/(1/2)\sigma_p^2$, so the regression coefficient of offspring on midparent equals the narrow-sense heritability.

The same reasoning can be followed to obtain covariances between other pairs of relatives, as summarized in Table 8. As can be seen, the additive genetic variance can be estimated directly either from parent-offspring covariance or from half-sib covariance. However, full-sib covariance includes a term resulting from dominance. The expressions in Table 8 are correct as long as there are no complications such as genotype-environment associations or other nonrandom environmental effects such as full sibs sharing environmental factors common to the whole family but not shared by other families. Since the total variance in phenotypic value σ_p^2 can be estimated

directly, once σ_a^2 is estimated from the covariance between relatives, the narrow-sense heritability can be estimated from Equation 8.31. The first three relationships in Table 8 are the most useful in quantitative genetics and are commonly used in animal and plant breeding. The other relationships are used mainly in human quantitative genetics.

The genetic covariance between various relatives can also be derived using the concepts of gene identity developed by Cotterman (1940) and extended by Crow and Kimura (1970). In these terms, the generalized covariance for a pair of related individuals is

$$\sigma_{xy} = r\sigma_a^2 + u\sigma_d^2 \tag{8.36}$$

where the coefficients r and u are determined from coefficients of coancestry defined in Chapter 5. The coefficient of coancestry, F_{XY}, of two individuals X and Y is the inbreeding coefficient of a hypothetical offspring of X and Y. If individuals A and B are the parents of X, and C and D are the parents of Y, then the r and u coefficients in Equation 8.36 are

$$\begin{aligned} r &= 2F_{XY} \\ u &= F_{AC}\,F_{BD} + F_{AD}\,F_{BC} \end{aligned} \tag{8.37}$$

It is worthwhile to write out pedigrees for some of the relations in Table 8, calculate the coefficients of coancestry, and verify that Equation 8.36 gives the correct result.

Figure 17 presents the narrow-sense heritabilities of diverse quantitative traits in farm animals and one important crop plant as estimated from the

Table 8. Theoretical covariance in phenotype between relatives.[a]

Degree of relationship	Covariance
Offspring and one parent	$\sigma_a^2/2$
Offspring and average of parents (midparent)	$\sigma_a^2/2$
Half siblings	$\sigma_a^2/4$
Full siblings	$(\sigma_a^2/2) + (\sigma_d^2/4)$
Monozygotic twins	$\sigma_a^2 + \sigma_d^2$
Nephew and uncle	$\sigma_a^2/4$
First cousins[b]	$\sigma_a^2/8$
Double first cousins	$(\sigma_a^2/4) + (\sigma_d^2/16)$

[a] Variance terms due to interaction between loci (epistasis) have been ignored.
[b] First cousins are the offspring of matings between siblings and unrelated individuals; double first cousins are the offspring of matings between siblings from two different families.

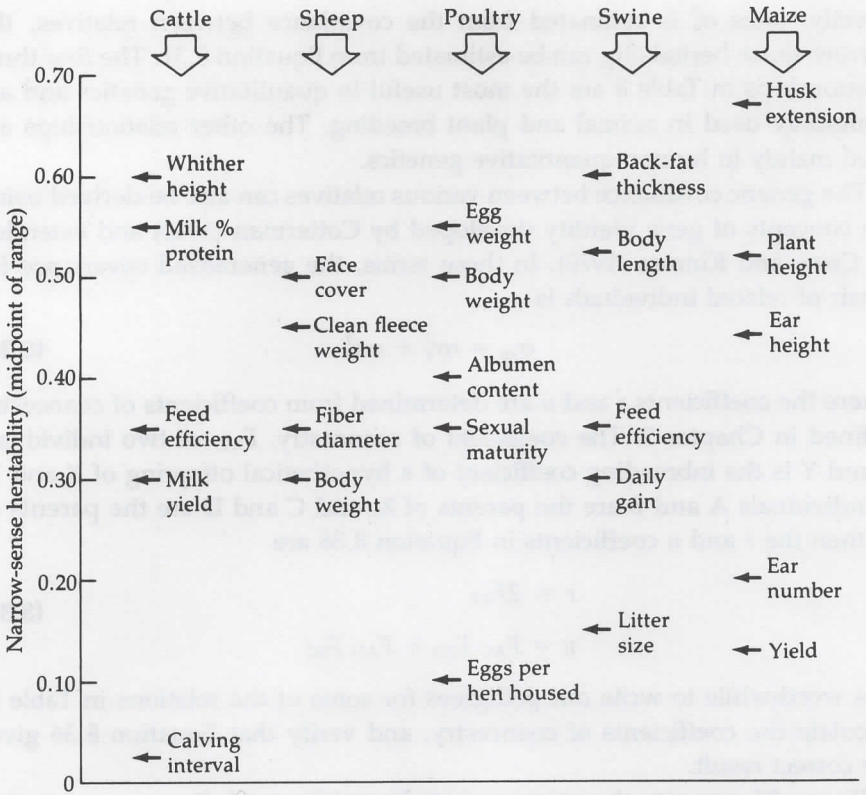


Figure 17. Narrow-sense heritabilities for representative traits in plants and animals. Traits closely related to fitness (calving interval, eggs per hen, litter size of swine, yield and ear number of corn) tend to have rather low heritabilities. (Animal data from Pirchner 1969, who gives the range of heritabilities in various studies. The midpoint of the range is plotted here. Corn data from Robinson et al. 1949.)

correlation between relatives. The data are presented merely to show the values of heritability with which breeders typically must deal. It is important to keep in mind that the heritabilities in Figure 17 pertain to one population in one type of environment at one particular time. The same trait in a different population or in a different environment might well have a different heritability. Generally speaking, traits that are closely related to fitness (such as calving interval in cattle or eggs per hen in poultry) tend to have rather low heritabilities. Ignoring complications such as antagonistic pleiotropy (discussed later), long-term natural selection is expected to gradually reduce the additive genetic variance until the effect is balanced against the input of new mutations.

For purposes of comparison, Figure 18 shows estimated broad-sense heritabilities of a number of quantitative traits in Caucasians. Broad-sense her-

itabilities vary widely for different traits, as they do in other species. Note the low heritability of fertility, a trait that is obviously closely related to fitness. The high heritabilities of IQ in these studies may be inflated by parent-offspring environmental covariance, an estimation problem that is almost impossible to solve in human studies.

Although it is tempting to think about resemblance between relatives in terms of classical genetic analysis such as Mendel did, there are major differences between the approaches. When the data are measurements of a continuous character in family-structured samples from a population, estimates of statistical components of variance can be obtained. However, these components depend on allele frequencies and environmental conditions, and thus quantities such as heritability are far removed from the basic level of gene action. Direct experimental assessment of variance components shows that different populations do have different heritabilities of many traits (e.g., Mitchell-Olds 1986). Moreover, a trait with high heritability does not mean that the trait cannot be affected by the environment. For example, the rare disease phenylketonuria is caused by a single defective gene for the enzyme phenylalanine hydroxylase, yet the phenotype of phenylketonuria—severe mental retardation—can be completely circumvented by a diet low in phenylalanine.

The distinction between estimation of variance components and knowledge of genetic causality for human traits has particularly important implications for social applicability of the findings, which are forcefully conveyed by Lewontin (1974b) and Feldman and Lewontin (1975). In particular, estimates of heritability within a population, even if they are sound, tell us nothing about the degree to which genetic differences account for the mean phenotypic differences between populations. Experimentalists working with

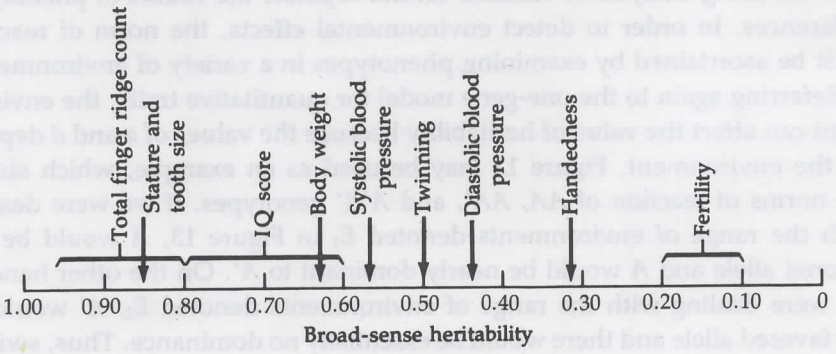


Figure 18. Broad-sense heritabilities and ranges of heritabilities of various traits in humans. Uncertainties about the correlation between environments of relatives make such estimates in humans very tentative. (Data from Smith 1975.)

organisms that can be manipulated in the field or laboratory can be more rigorous in the assessment of genetic parameters by examining the traits in several environments, and by doing studies that combine classical and quantitative genetic analysis.

NORM OF REACTION AND PHENOTYPIC PLASTICITY

In considering the influence of environment on the determination of phenotypes, it is tempting to think of environmental effects as random noise added to traits that are basically genetically determined. This view is grossly misleading. One simple way to analyze the effect of environment on phenotype is to examine the phenotypes of a single genotype in an array of environments, which is possible to do in a number of experimental organisms. The array of phenotypes that results from a given genotype is known as the **norm of reaction**, a term coined by Schmalhausen (1949). Figure 19 shows the norm of reaction of bristle number for two strains of *Drosophila pseudoobscura* reared at three temperatures (Gupta and Lewontin 1982). Note that the two strains have quite different responses to temperature. Comparison of panels a and b reveals an important feature of quantitative traits and environmental effects. If we were to do a quantitative genetic study of parent–offspring regression or some other means of determining components of variance, with all the experiments done at a temperature of about 19°C (panel a), there would be marked differences seen among the genotypes, and the heritability would consequently be high. At 24°C (panel b), the norms of reaction converge, the phenotypes are much more similar, and the heritability would be low. This contrast underscores the fact that heritability is a measure defined in one environment, and it shows that the methods of quantitative genetics using analysis of variance cannot separate the causes of phenotypic differences. In order to detect environmental effects, the norm of reaction must be ascertained by examining phenotypes in a variety of environments.

Referring again to the one-gene model for quantitative traits, the environment can affect the value of heritability because the values of a and d depend on the environment. Figure 13 may be used as an example, which shows the norms of reaction of AA, AA', and $A'A'$ genotypes. If we were dealing with the range of environments denoted E_1 in Figure 13, A would be the favored allele and A would be nearly dominant to A'. On the other hand, if we were dealing with the range of environments denoted E_2, A' would be the favored allele and there would be essentially no dominance. Thus, switching a population from E_1 to E_2 would change the values of a and d and substantially alter the heritability of the trait, even though the total phenotypic variance of the population might remain the same.

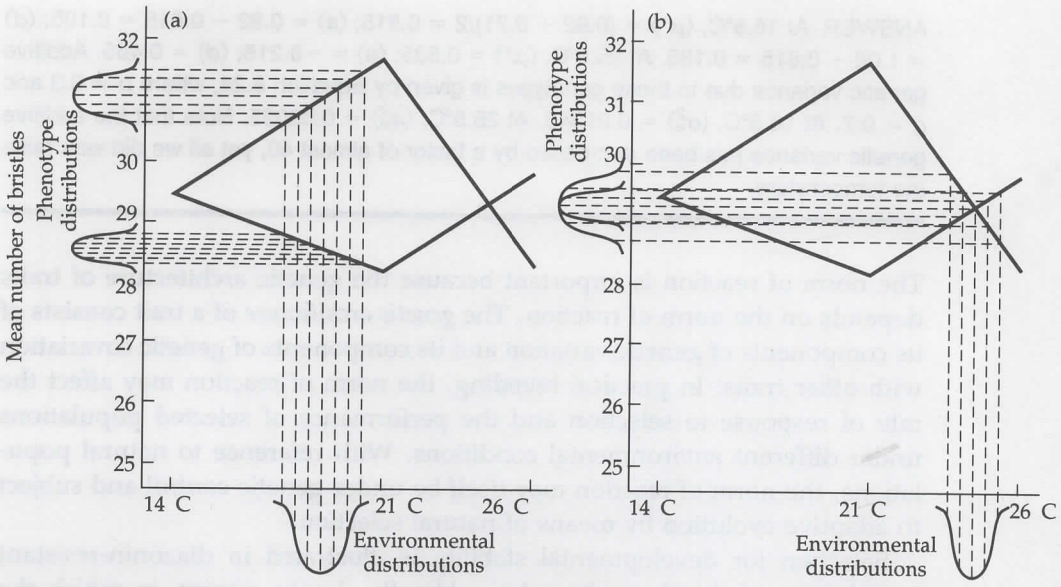


Figure 19. The norms of reaction of bristle number for *Drosophila pseudoobscura*. The plots show the environmental effect on genetic variance and the impact of genotypic distribution on environmental variance using two actual genotypic norms of reaction. (a) At temperatures from 17 to 21°C, the two genotypes have very different phenotypic distributions. (b) At higher temperatures, averaging about 23°C, the phenotypic distributions are nearly coincident. (From Gupta and Lewontin 1982.)

Norm of Reaction

F

Dobzhansky and Spassky (1944) studied the norms of reaction with respect to temperature and viability of various genotypes containing chromosomes extracted from natural populations of *Drosophila pseudoobscura*. For two such chromosomes (designated A and B), the following relative viabilities were obtained.

Temperature (°C)	Genotype		
	A/A	*A/B*	*B/B*
16.5	0.92	1.00	0.71
25.5	0.32	1.00	0.75

Estimate μ^*, a, and d for these genotypes in populations maintained at 16.5 and 25.5°C. Then, letting p represent the frequency of the A chromosome and q that of the B chromosome, assume $p = 0.3$, and estimate the additive genetic variance of viability resulting from these genotypes at both temperatures.

ANSWER At 16.5°C, $\langle \mu^* \rangle = (0.92 + 0.71)/2 = 0.815$; $\langle a \rangle = 0.92 - 0.815 = 0.105$; $\langle d \rangle = 1.00 - 0.815 = 0.185$. At 25.5°C, $\langle \mu^* \rangle = 0.535$; $\langle a \rangle = -0.215$; $\langle d \rangle = 0.465$. Additive genetic variance due to these genotypes is given by Equation 8.24, where $p = 0.3$ and $q = 0.7$. At 16.5°C, $\langle \sigma_a^2 \rangle = 0.01346$. At 25.5°C, $\langle \sigma_a^2 \rangle = 0.00035$. Note that the additive genetic variance has been decreased by a factor of almost 40, yet all we did was raise the temperature.

The norm of reaction is important because the genetic architecture of traits depends on the norm of reaction. The *genetic architecture* of a trait consists of its components of genetic variation and its components of genetic covariation with other traits. In practical breeding, the norm of reaction may affect the rate of response to selection and the performance of selected populations under different environmental conditions. With reference to natural populations, the norm of reaction may itself be under genetic control and subject to adaptive evolution by means of natural selection.

Selection for developmental stability is illustrated in diazonin-resistant populations of the Australian sheep blowfly, *Lucilia cuprina*, in which the insecticide resistance results from a single dominant gene. In the genetic background of the evolved natural population, the resistant genotypes show less sensitivity in their developmental response to environmental fluctuations than do the progeny of crosses in which the genetic background has been replaced with that of an unselected laboratory population. This is reflected in the finding of a smaller average degree of asymmetry between the left and right sides of the body (measured by the difference in bristle number between the two sides) in the natural population (Clarke and McKenzie 1987; Jones 1987). The implication of this finding is that natural selection can favor genetic modifiers that influence the manner in which genotypes respond to the environment.

As is evident in Box F, the heritability of a trait has reference only to a particular population at a particular time in a particular range of environments. Because practical breeders are usually interested in improving a particular population at a particular time in a particular range of environments, the concept of heritability is of great utility when used for its intended purpose—the prediction of response to selection.

Another factor that affects the correspondence between genotypes and phenotypes is age. A phenotype, such as body weight, often changes with age, and different genotypes can have different age-related growth curves (or other developmental profiles). Consequently, the heritability of traits depends on the age at which individuals are tested. Figure 20 shows components of variance of mouse body weight at different ages (Riska et al. 1984). In this study, 2700 mice from 700 full-sib families were measured from days 14 to 70. Variance of all components was maximal at an age of about

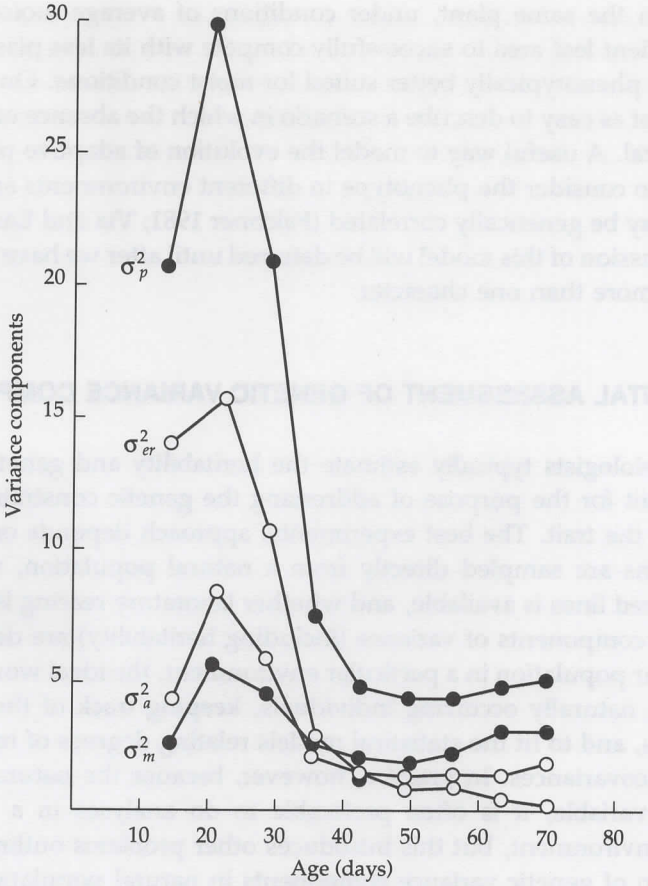


Figure 20. Variance components for the logarithm of body weight in a randombred (genetically variable) strain of mice plotted as a function of age. σ^2_m is the variance due to maternal effects, and $\sigma^2_{er} = \sigma^2_e - \sigma^2_m$. (From Atchley 1984.)

20 days (which happens to be the time of maximal growth), and then declined to fairly stable values at age 40 days. Despite the additive genetic variance decreasing markedly after 20 days of age, the heritability itself actually increased with age, with values of 0.22, 0.27, 0.31, and 0.37 at days 12, 20, 50, and 70. The tendency for increasing heritability with age is seen in many traits in humans as well.

The degree to which the environment shapes the phenotype has also been called **phenotypic plasticity**. It is interesting to consider whether natural selection modifies phenotypic plasticity as part of adaptation, especially since extreme phenotypic plasticity is not always favorable. For example, if a plant germinates from a seed in an exceptionally dry year, because its genotype has a high level of phenotypic plasticity that enables it to cope well with dry

periods, then the same plant, under conditions of average moisture, may have insufficient leaf area to successfully compete with its less plastic neighbors that are phenotypically better suited for moist conditions. On the other hand, it is just as easy to describe a scenario in which the absence of plasticity can prove fatal. A useful way to model the evolution of adaptive phenotypic plasticity is to consider the phenotype in different environments as different traits that may be genetically correlated (Falconer 1981; Via and Lande 1985). Further discussion of this model will be deferred until after we have discussed selection of more than one character.

EXPERIMENTAL ASSESSMENT OF GENETIC VARIANCE COMPONENTS

Population biologists typically estimate the heritability and genetic correlations of a trait for the purpose of addressing the genetic constraints on the evolution of the trait. The best experimental approach depends on whether the organisms are sampled directly from a natural population, whether a series of inbred lines is available, and whether laboratory rearing is practical. Because the components of variance (including heritability) are descriptions of a particular population in a particular environment, the ideal would appear to be to use naturally occurring individuals, keeping track of their familial relationships, and to fit the statistical models relating degrees of relationship to expected covariances. In practice, however, because the natural environment is so variable, it is often preferable to do analyses in a controlled laboratory environment, but this introduces other problems outlined below.

Estimation of genetic variance components in natural populations generally requires a number of restrictive assumptions: (1) disomic inheritance, (2) no correlation of parental and offspring environments, (3) no linkage or linkage disequilibrium, (4) parents equally inbred, (5) offspring not inbred, (6) samples of relatives drawn at random from a noninbred population, (7) no mutation, migration, selection, and (8) random mating (see Mitchell-Olds and Rutledge 1986 and references therein). In practice, small violations of these assumptions are acceptable, but it is nevertheless a serious challenge to obtain reliable estimates of genetic variance components from samples from natural populations. Most commonly a full analysis is not carried out in natural populations, but offspring collected from known mothers are studied. An example of this approach is presented in Box G.

As Box G indicates, once measurements are obtained on individuals with known degrees of relationship, the partitioning of variance into additive, dominance, and environmental components can be done using the standard statistical method of analysis of variance. Many experimental designs permit the effects of various factors on quantitative genetic components to be esti-

mated. For example, the experiment in Box G could be repeated under several different environmental regimens, and the environmental component in the analysis of variance could be estimated. A second generation of plants could also be studied and the variance components estimated from parent-offspring covariance. However, the two most common designs are analysis of variance of full-sib families or half-sib families. Using full-sib families alone has the problem that the covariance of full sibs includes both additive and dominance variance. Therefore, when the data are limited to full-sib families, all that can be estimated is the broad-sense heritability. On the other hand, if one has only half-sibs [whose covariance is $(1/4)\sigma_a^2$], one can estimate the narrow-sense heritability, but the dominance component may be quite large and remain undetected.

Components of Genetic Variance in Natural Populations

G

The study of Mazer (1987) typifies the approach of sampling offspring from a natural population and performing a half-sib analysis of variance. The character of interest was seed weight, because the wild radish *Raphanus raphanistrum* shows a 20-fold range in seed weight. The species is an obligate outcrosser, so seeds from a single female plant must be related as either full sibs or half sibs. In the study, groups of half-sib progeny obtained from hand pollination in a greenhouse were grown outdoors and studied for many phenotypic attributes, including stem number, stature, lifetime flower and fruit production, seed number and seed weight. Partitioning of variance components was done using a widely available computer statistical package that performs analysis of variance. The factors in the analysis of variance included a maternal effect, a paternal effect, an interaction term, and an error term. The computer program partitions the variance into such components as σ_s^2 (the "sire" or paternal component of variation) and σ_d^2 (the "dam" or maternal component of variation). The narrow-sense heritability for the half-sib design equals $4\sigma_s^2/\sigma_p^2$, and was estimated as 0.08 ± 0.97. Hence we cannot reject the null hypothesis that the additive genetic variance for seed weight is zero in this population. This means that, despite the great range in seed weight, the trait may not respond to individual selection, because the variation may be due solely to a combination of maternal effects, nonadditive genetic effects, and environmental effects. More elaborate designs would allow a more complete partitioning of these components.

More elaborate designs that feature parents, full sibs, and half sibs can give extensive partitioning of variance components. The reliability of the methods can also be tested by comparing heritability estimates from either parent-offspring regression or from the covariance of half sibs. One means of using all the data in a single estimate is the method of **maximum likelihood**, a procedure that we encountered first in Chapter 1 for allele frequency

estimation. Maximum likelihood methods are preferable to analysis of variance when several traits are being examined in order to assess components of genetic covariance. Analysis of variance becomes extremely cumbersome when sample sizes are not the same at all levels of relationship, and there is really no single ideal way to adjust for unequal sample sizes.

The principle behind maximum likelihood is to construct a likelihood function that describes the likelihood of obtaining the observed data given the family structure and a set of unknown parameters to be estimated. The unknown parameters are the magnitudes of the various genetic and environmental variance components. The method then finds the values of the unknown parameters that maximize the likelihood. In practice, the unknown parameters form a variance-covariance matrix, and the computer algorithms entail extensive matrix manipulations. The computer output consists of estimates of heritability that utilize all the data, including parent-offspring, full sib, half sib, and any other relationships that are informative. Although no assumption of multivariate normality of the character data is necessary in obtaining the estimates, the estimates only have meaning (i.e. the prediction equation is only true) if the phenotypes are normally distributed, and testing the statistical significance of the estimates requires normality. Shaw (1987) provides a thorough review of the merits of maximum likelihood methods in quantitative genetics.

Heritability says virtually nothing about the actual mode of inheritance of a quantitative trait, useful as the concept may be in predicting response to selection. The heritability of a trait represents the cumulative effect of all genes that affect the trait. Even for a single gene, heritability depends in a complex manner on the values of p, a, and d, and these individual components cannot be disentangled. (The values of p, a, and d are said to be statistically *confounded*.) With more than one gene, the heritability includes a summation of terms for each gene, and each term has its own particular values of p, q, a, and d. Here, precisely, is the problem: for a quantitative trait determined by, say, 10 diallelic loci, there would be 30 quantities involved in heritability—10 allele frequencies, 10 values of a, and 10 values of d. Heritability is but a single number that gives the combined effect of all 30 quantities. It says nothing about any one of them. With more genes affecting the trait or with multiple alleles, matters would be even worse. Also, we have assumed that all genes affecting the trait act independently of one another and are unlinked. In actual cases, genes often interact and can be linked. Moreover, the assumption of no correlation of parental and offspring environments is not always easy to test. All in all, while heritability, especially realized heritability, is an indispensable aid to plant and animal breeders, it lends itself to no easy interpretation in simple genetic terms (apart from the statistical description through parent-offspring regression). Another difficulty in interpreting heritability values is that they depend on the range of envi-

ronments that occur. The denominator σ^2 ($= \sigma_P^2$) in Equation 8.23 is the total variance in phenotypic value in the population. Because the total variance includes the variance resulting from environmental differences among individuals, increasing the variation in the environment decreases h^2. Exceptionally thoughtful discussions of the concept of heritability and its strengths and limitations are found in Kempthorne (1978) and Jacquard (1983).

Heritability values are determined in part by gene frequencies. Because gene frequencies change during the course of selection, the heritability is also expected to change. In practice, however, the heritability changes sufficiently slowly that over the course of a few generations it can be regarded as a constant.

The approximate constancy of heritability has a twofold cause: (1) if a particular gene accounts for only a small proportion of the total phenotypic variance in a quantitative trait, then the gene frequency does not change very rapidly, and (2) the values of a and d remain nearly constant provided that the environment does not change drastically from one generation to the next. Thus, at least for the first ten generations or so, heritability usually remains approximately constant and can be used as a constant in the prediction equation (Equation 8.10). To be precise, suppose h^2 is constant and let μ_t and S_t represent the mean of the population and the selection differential in the t^{th} generation. Then, over the length of time during which h^2 is approximately constant,

$$\mu_t - \mu_0 = h^2(S_0 + S_1 + \ldots + S_{t-1}) \qquad (8.38)$$

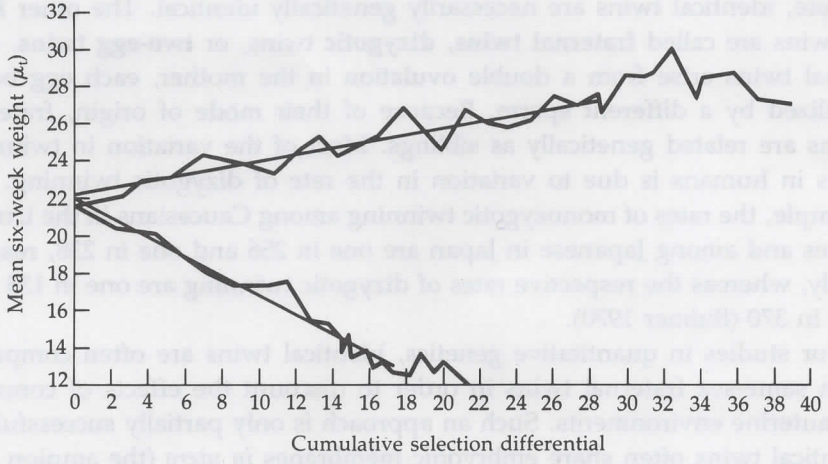


Figure 21. Linearity in response against the cumulative selection differential for body weight in mice at age six weeks. Linearity in the up (high weight) direction continues for about twice as long as it does in the down (low weight) direction. (After Falconer 1955.)

The quantity $\mu_t - \mu_0$ is the total response to selection, and $S_0 + S_1 + \ldots + S_{t-1}$ is called the **cumulative selection differential**. During the time in which h^2 is approximately constant, therefore, a plot of μ_t against the cumulative selection differential is expected to yield a straight line with slope equal to h^2, as illustrated for a case in mice in Figure 21.

TWIN STUDIES AND INFERENCES OF HERITABILITY IN HUMANS

Identical twins are among the favorite degrees of relationship for study in human genetics. Because identical twins are genetically identical, phenotypic differences between identical twins would seem to be a straightforward reflection of how much phenotypic variance is caused by environment. Twin studies raise their own unique problems, however, and the results must be interpreted with caution. Before discussing the use of twins in quantitative genetics, we should back up a few steps and first discuss the phenomenon of twinning itself.

Twins are relatively frequent among human births, though the rate of twinning varies from population to population. Among Caucasians in the United States, for example, about one in 88 births results in twins; among Japanese in Japan, the rate is about one in 145 births (Bulmer 1970). Two kinds of twins actually occur. **Identical twins**, often called **monozygotic** or **one-egg** twins, arise from a single zygote that very early in embryonic development splits into two distinct clumps of cells, each clump thereafter undergoing its own embryonic development. Because they arise from a single zygote, identical twins are necessarily genetically identical. The other kind of twins are called **fraternal twins**, **dizygotic** twins, or **two-egg twins**. Fraternal twins arise from a double ovulation in the mother, each egg being fertilized by a different sperm. Because of their mode of origin, fraternal twins are related genetically as siblings. Most of the variation in twinning rates in humans is due to variation in the rate of dizygotic twinning. For example, the rates of monozygotic twinning among Caucasians in the United States and among Japanese in Japan are one in 256 and one in 238, respectively, whereas the respective rates of dizygotic twinning are one in 135 and one in 370 (Bulmer 1970).

For studies in quantitative genetics, identical twins are often compared with same-sex fraternal twins in order to discount the effects of common intrauterine environments. Such an approach is only partially successful, as identical twins often share embryonic membranes *in utero* (the amnion and chorion) that are not usually shared by fraternal twins. Moreover, because identical twins often have astonishingly similar facial features, they may be treated more similarly by parents, teachers, and peers than are fraternal

twins. Some of these problems can be overcome by studying twins that are raised apart (in different households), but data of this sort are limited (Shields 1962), making estimates of heritability highly imprecise. Even when twins are reared apart, the environments into which they are adopted are generally similar. This effect of correlated environments inflates the apparent degree to which traits are genetically determined. In any case if r_{MZ} and r_{DZ} represent the correlation coefficients of a quantitative trait among monozygotic and dizygotic twins, then $2(r_{MZ} - r_{DZ})$ provides a rough estimate of the broad-sense heritability of the trait. [See Smith 1975, for a good discussion. Recall that the correlation coefficient, r, in a set of pairs of numbers (x, y) is defined as $r = \sigma_{xy}/\sigma_x\sigma_y$, where σ_{xy} is the covariance between x and y and where σ_x and σ_y represent the standard deviations of x and y.]

THRESHOLD TRAITS AND THE GENETICS OF LIABILITY

Some multifactorial traits do not exhibit continuous variation. Although the variation is discontinuous (individuals either express the trait or not), the trait is nevertheless influenced by multiple genetic factors and also by environment. Such traits are called **threshold traits**. A human example is diabetes, an abnormality in sugar metabolism that affects 1% or 2% of the Caucasian population. In a sense, diabetes is a continuous trait because the severity of the disease varies from nearly undetectable to extremely severe. On the other hand, diabetes can also be considered a threshold trait because all individuals may be classified according to whether or not they are so severely affected that clinical treatment is required. With such a classification, there are only two phenotypes, "affected" and "not affected," even though there is phenotypic variation within each category. The genetic influence on the trait is shown by the enhanced risk of diabetes in relatives of affected individuals. However, environmental factors such as diet are also important in determining whether high-risk genotypes actually develop the disease. At one time, many threshold traits were "explained" by postulating a simple genetic mechanism (such as a single recessive allele in the case of diabetes) and invoking "incomplete penetrance" to account for the poor fit of pedigree data to a simple Mendelian hypothesis. Nowadays it is preferred, and probably more realistic, to consider threshold traits as bona fide polygenic traits and to calculate heritabilities as for any other quantitative trait. In most cases, however, it is simply not known whether the genetic influence on the trait results from one or a few major genes or is polygenic.

The basic idea behind the model of threshold traits is illustrated in Figure 22. The normal curve in panel (a) represents the (unobservable) distribution of a hypothetical liability (or risk) toward the threshold trait, measured on a

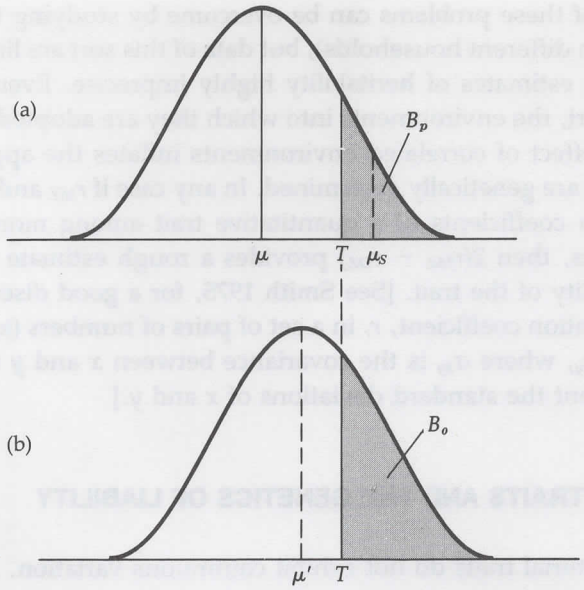


Figure 22. (a) Distribution of liability assumed for a threshold trait in a hypothetical population. The shaded area denotes individuals having liability above a critical threshold (T) and consequently affected with the trait. B_p is the frequency of affected individuals in the entire population, μ is the mean liability of individuals in the entire population, and μ_s the mean liability of affected individuals. (b) Distribution of liability among offspring who have one parent affected with the trait. μ' denotes the mean liability among offspring, and B_o is the proportion of affected offspring. Note the similarity to Figure 6. In Figure 6, however, the quantitative trait could be measured directly in each individual. In this case, liability cannot be measured directly, and inferences about the distribution of liability must be made from B_p and B_o.

scale such that the mean value is 0 and the variance is 1. It is assumed that individuals whose liability is above a certain threshold (T) actually express the trait. Thus, the shaded area in Figure 22a delimits the proportion of individuals in the population who are affected (B_p), and the mean liability among affected individuals is denoted μ_s. Figure 22b gives the (again unobservable) distribution of liability among the offspring of affected individuals. The offspring mean is denoted μ', and the proportion of offspring above the threshold is denoted B_o. The setup here is like that in the earlier section of this chapter in which we calculated the regression coefficient of offspring on one parent. In this model it can be shown that the regression coefficient b is given by $b = \mu'/\mu_s$, and the appropriate estimate of the heritability of liability is obtained from the relation

$$h^2 = 2b = 2\mu'/\mu_s \qquad \textbf{(8.39)}$$

To illustrate the calculations, we will use data from Carter (1961) on the incidence of pyloric stenosis, which is an obstruction of the opening at the lower end of the stomach. The incidence of the condition differs in males and females, but we will use the incidence in fathers and their sons. For pyloric stenosis, the incidence among males in the general population is $B_p = 0.005$, and the incidence among sons of affected males is $B_o = 0.05$. From these two numbers, we wish to infer μ_s and μ' in order to calculate the heritability of liability.

The mean liability of fathers (μ_s) can be obtained immediately from Figure 23, which gives the relation between the **intensity of selection** i (where $i = S/\sigma$) and the proportion of the population B to the right of the truncation point, along with convenient approximations. For $B_p = 0.005$, Figure 23 gives $i = 2.89$. Since $S = \mu_s - \mu = i\sigma$ and we have assumed that $\mu = 0$ and $\sigma^2 = 1$, we have it that $\mu_s = 2.89$ also. The quantity B_p can also be used to obtain the threshold (T) using tables of the normal distribution. Note that B_p rep-

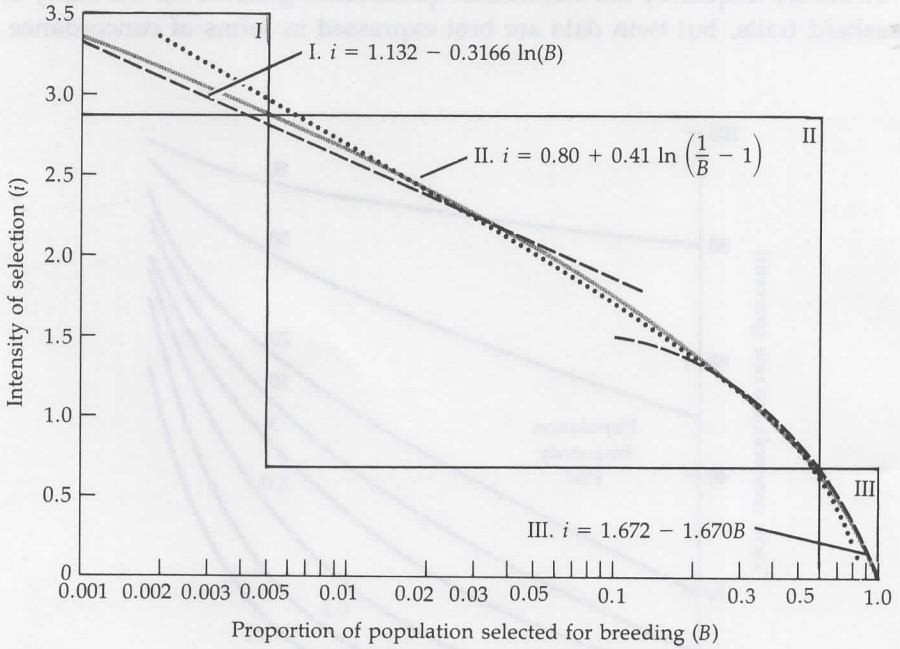


Figure 23. The gray curve is the relationship between the intensity of selection (i) and the proportion of the population saved for breeding (B). Note that the scale of B is logarithmic. The dashed and dotted lines are approximations to the actual curve devised by Simmonds (1977). Approximation I is suitable for the range $B = 0.001$ to 0.005 or $i = 3.4$ to 2.9; approximation II for the range $B = 0.005$ to 0.6 or $i = 2.9$ to 0.65; and approximation III for the range $B = 0.6$ to 1 or $i = 0.65$ to 0. These approximations are sufficiently accurate for most purposes.

resents individuals with a positive deviation from the mean of an amount T or greater. Because of the symmetry of the distribution, B_p also represents individuals with a negative deviation of T or greater (although individuals with negative deviations are not affected). Overall, $2B_p$ is the proportion of the population having a positive or negative deviation of T or more units, and for $2B_p = 0.010$, the tabulated value of $T = 2.58$.

From the value of T above and the observed value of B_o, the mean liability of the offspring μ' can be obtained from the Gaussian distribution. Note in Figure 22b that the offspring distribution can be slid μ' units to the left to coincide with the parental distribution. The sliding procedure shifts the offspring threshold to the value $T - \mu'$. Among the offspring, therefore, $2B_o$ represents the proportion of the population having a positive or negative deviation of $T - \mu'$ or greater, and this gives $T - \mu' = 1.64$ for $2B_o = 0.10$. From the paragraph above, we know that $T = 2.58$, so $\mu' = 2.58 - 1.64 = 0.94$. Finally, the estimate of heritability of liability of pyloric stenosis based on father-son incidences is $\langle h^2 \rangle = 2\langle \mu' \rangle / \langle \mu_s \rangle = 2(0.94)/(2.89) = 65\%$.

Twins are frequently used in human quantitative genetics for the study of threshold traits, but twin data are best expressed in terms of concordance.

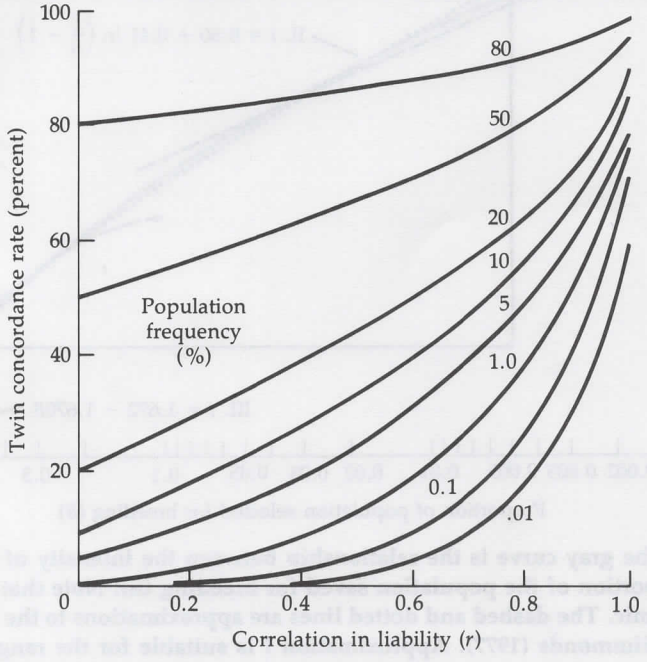


Figure 24. Threshold concordance rates expected in monozygotic twins, plotted against the correlation in liability toward the trait and the population incidence of the trait. (From Smith 1975.)

The **concordance** of a trait in a population of twins is the proportion of affected twins that have affected co-twins. For example, suppose that 100 affected individuals are found to be twins and that in 35 cases the co-twin is also affected. The concordance rate is then 35/100 = 35%. From the concordance rates for monozygotic and dizygotic twins and the incidence of the trait in the population, the correlations in liability between monozygotic twins (r_{MZ}) and dizygotic twins (r_{DZ}) can be calculated (Figure 24). The broad-sense heritability is then estimated by $2(r_{MZ} - r_{DZ})$, as discussed earlier. As with quantitative traits in general, twin data are more reliable if the twins are reared apart, but it is seldom possible in practice to obtain a sufficient number of such twin pairs.

Estimation of the Number of Genes Affecting a Quantitative Character

The number of genes that contribute to quantitative traits is not always large. We have already seen an example of seed color in wheat in which the number of genes was three. When the number of genes is relatively small, the number can often be estimated from the means and variances observed in different strains and their hybrids and backcrosses. Figure 25 gives an example with two alleles of each of three genes, which, as in the wheat example, are additive in their effects. The solid dots represent alleles that contribute one unit to phenotype and the open circles represent alleles that contribute nothing. The parental strains are homozygous for all unfavorable or all favorable alleles, and they differ by six units in phenotypic value. The variance in phenotypic value in the F_2 generation equals 3/2 units^2. Extending the model for two additive alleles of each of n unlinked genes, the difference D in phenotypic value between the means of parental inbred lines is $2n$ units, and the variance σ^2 in the F_2 generation is $n/2$ units^2.

In order to obtain an estimate of the number of genes that is independent of the units of measurement, a ratio is needed to make the units cancel. One possibility, first suggested by Wright (in Castle 1921), is

$$\langle n \rangle = \frac{D^2}{8\sigma^2} \qquad \textbf{(8.40)}$$

With n unlinked, additive genes, $\langle n \rangle = (2n)^2/[8(n/2)] = n$, as it should. Equation 8.40 is based on the assumptions of complete additivity, equal effects of all genes, no linkage, and fixed differences between parental lines. When the assumptions are violated, application of the equation usually results in estimates of gene number that are smaller than the actual numbers. For this reason, the quantity estimated in Equation 8.40 is called the **effective number** of genes because it defines a lower limit to the actual number.

The variance in Equation 8.40 is the **genetic variance**, which is the variance

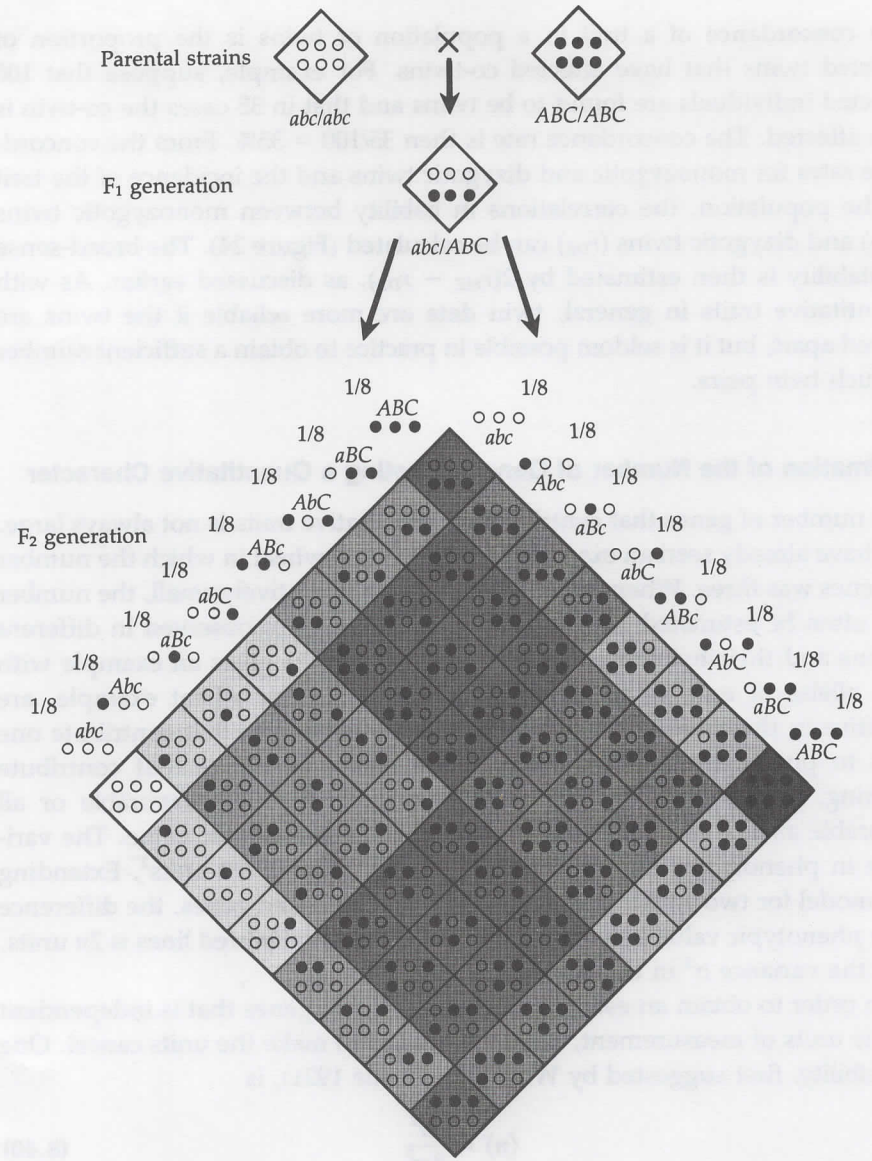


Figure 25. Segregation of three unlinked, additive genes affecting a quantitative trait. Each uppercase allele in a genotype contributes one unit to the phenotypic value. In the F_2 generation, the mean phenotypic value equals 3.0 and the variance is 1.5.

in phenotypic value resulting from genetic differences among individuals. When the environment contributes an amount σ_e^2 to the phenotypic variance, then the variance within the parental inbred lines, or within the F_1 population, equals σ_e^2. This is because the populations are genetically uniform, and

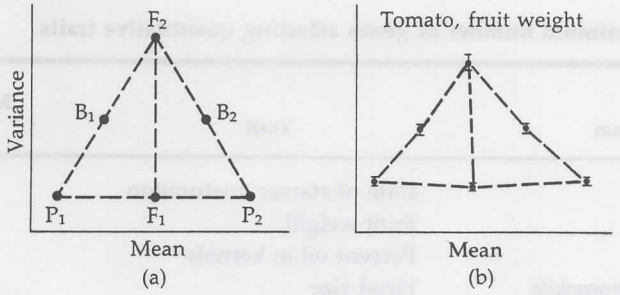


Figure 26. (a) Expected triangular relation between means and variances of phenotypic value among inbred parents (P), backcross progeny (B), and hybrids (F) for an ideal quantitative trait determined by unlinked and completely additive genes. (b) Observed relation for fruit weight in tomato, in which the index of fruit weight is *phenotypic value* = log[(*weight in grams*) − 0.153]. (After Lande 1981.)

the only source of variation in phenotype results from the environment. However, within the F_2 generation, variation in phenotype results in part from genetic variation and in part from environmental variation, and the total variance in phenotypic value equals the summation $\sigma^2 + \sigma_e^2$. Therefore, subtraction of the F_1 variance from the F_2 variance gives an estimate of σ^2 because

$$\sigma^2 = (\sigma^2 + \sigma_e^2) - \sigma_e^2 \qquad (8.41)$$

Further discussion of the genetic variance occurs later in this chapter, and Lande (1981) gives several alternative methods of estimating the genetic variance using data from inbreds, hybrids and backcrosses. Cockerham (1986) extended the analysis to obtain an unbiased estimation of the difference in parental means, and he combined the data from parentals, F_1, F_2, and backcrosses into a single least-squares estimate.

The scale in which phenotype is measured is important in using Equation 8.40 because the genes and alleles are assumed to be additive. In the ideal additive case, a plot of the means and variances of inbreds, hybrids and backcrosses forms a triangle of the type shown in Figure 26(a). In actual cases, the phenotypes must be measured using a scale which yields an approximate triangle. With fruit in tomatoes, for example, an obvious scale of weight is in grams, but the scale giving the triangle in Figure 26(b) is x = log[(weight in grams) − 0.153] (Lande 1981). Using this scale, $D = 1.552$ and $\sigma^2 = 0.0426$, from which Equation 8.40 implies $\langle n \rangle = (1.552)^2/[8(0.043)] = 7$ as the effective number of genes.

The effective number of genes for oil content in maize kernels calculated in Box H bracket those obtained by Lande (1981) using other estimates of the environmental variance. A summary of the estimated number of genes affecting various quantitative traits is given in Table 9. With the notable

Table 9. Minimum number of genes affecting quantitative traits.

Organism	Trait	Minimum number of genes
Goldenrod	Date of stamen maturation	6–7
Tomato	Fruit weight	7–11
Maize	Percent oil in kernels	17–22
Hawaiian *Drosophila*	Head size	6–9
Fish	Eye diameter	5–7
Laboratory rat	Coat coloration	5–8
Human	Skin color	4–6

(Data from Lande 1981.)

exception of oil content in maize kernels, the effective number of genes is less than ten. These examples demonstrate that genetic variation in quantitative traits can result from the segregation of a relatively small number of genes. However, the effective number of genes represents a minimum estimate, and the actual number of genes is likely to be larger. In some cases the estimated number of genes is very large. For example, at least 40 genes contribute antigens that are important in the rejection of skin transplants in mice, at least 150 genes contribute to body weight in the mouse and *Tribolium*, and at least 17 genes on the third chromosome of *Drosophila melanogaster* influence sternopleural bristle number (Shrimpton and Robertson 1988). It should be emphasized that the number of genes that affect quantitative traits may differ according to the precise definition of *affect*. For example, the overall long-term selection response illustrated in Figure 5 may result from a large number of genes, but most of the response that occurs at any one time may result from changes in allele frequency in only a few of them.

H Estimating the Number of Genes Determining Corn Oil Content

In analyzing data on oil content in maize, Lande (1981) found that values of log[(percent oil in kernels) + 1.87] approximated a triangular form like those in Figure 26. Using this scale of measurement, the means of two inbred lines were 0.513 and 1.122, respectively, and the phenotypic variances of the inbred lines, the F_1 generation, and the F_2 generation were 0.0014, 0.0005, 0.0003, and 0.0030, respectively. Use Equations 8.40 and 8.41 to estimate the effective number of genes. Estimate the environmental variance in two different ways: (1) as the mean of the variance in phenotypic value in the parental inbred lines, and (2) as equal to the variance in phenotypic value in the F_1 generation.

ANSWER Using the first estimate, $\langle \sigma_\varepsilon^2 \rangle = (0.00142 + 0.00053)/2 = 0.00098$, $\langle \sigma^2 \rangle = 0.00303 - 0.00098 = 0.00205$, $D = 0.513 - 1.122 = -0.609$, and $\langle n \rangle = (-0.609)^2/[8(0.00205)] = 22.6$. Using the second estimate, $\langle \sigma^2 \rangle = 0.00303 - 0.00030 = 0.00273$, and $\langle n \rangle = 17.0$.

CORRELATED RESPONSE AND GENETIC CORRELATION

Genes have pleiotropic effects on phenotype; that is, every gene potentially affects every trait in the organism, either as a primary effect or as a secondary, indirect effect. Therefore, the alleles that are favorable for one quantitative trait may have unfavorable effects on another quantitative trait, and as these alleles are increased in frequency by artificial selection (thereby improving the phenotypic value with respect to the selected quantitative trait), the very same alleles may bring about a deterioration of some other aspect of performance. Pleiotropy is one cause of **correlated response**—a change in phenotypic value of one trait that accompanies response to selection of a different trait. A second possible cause of correlated responses is linkage disequilibrium (Chapter 1)—a favorable allele for one trait that increases in frequency under selection may drag along with it an allele of another, tightly linked gene that has a detrimental effect on an unselected trait.

Correlated responses are quite common in artificial selection and often, but not always, result in a deterioration in reproductive performance. In the case of Leghorn chickens, for example, 12 generations of selection for increased shank length reduced the egg hatchability by nearly half (Lerner 1958). In turkeys, to take another example, there was intense selection during the period 1944–1964 for growth rate, body conformation, and body size, but there was also a steady decline in some aspects of reproductive fitness such as fertility, egg production, and egg hatchability (Nordskog and Giesbrecht 1964). On the other hand, correlated responses can sometimes be useful. For example, selection for larger mature body size often increases litter size in mice and swine. If a trait has a low heritability or is difficult to measure, it is sometimes possible to practice selection for another, correlated trait, obtaining progress in the trait of interest by correlated response. Theoretically, the maximum response to artificial selection occurs when the criterion for selection is determined by a *selection index* (an averaging across several traits) which takes genetic correlations into account. However, the theoretical advantage of index selection is often overridden by practical difficulties in estimating the components of the index and implementing the selection procedure.

From a theoretical point of view, the covariance between two quantitative traits can be partitioned in a manner analogous to the partitioning of the variance for one trait outlined in an earlier section; the covariance can thus be partitioned into an additive covariance, a dominance covariance, an environmental covariance, and so on. The most important theoretical result is that the amount of correlated response with individual selection depends only on the additive covariance, much as the direct response to individual selection depends only on the additive variance. The components of covariance between traits can be estimated from the resemblance between relatives, but often it is preferable to estimate the correlated response by direct observation in a manner analogous to the determination of realized heritability (Falconer 1981).

The phenotypic correlation is the correlation one would obtain by measuring two traits, say X and Y, and calculating the correlation coefficient directly. In symbolic terms, the phenotypic correlation is

$$r_p = \frac{\text{Cov}_p}{\sigma_{pX}\,\sigma_{pY}} \tag{8.42}$$

where Cov_p is the phenotypic covariance and σ_{pX} and σ_{py} are the phenotypic standard deviations of characters X and Y. Correlations do not partition in the same way as variances, but the phenotypic covariance can be expressed as the sum

$$\text{Cov}_p = \text{Cov}_a + \text{Cov}_e \tag{8.43}$$

where Cov_a is the additive genetic covariance, and Cov_e is the environmental covariance. The genetic correlation (which is essentially the correlation of additive genetic effects) is defined as

$$r_a = \frac{\text{Cov}_a}{\sqrt{(\sigma^2_{aX}\,\sigma^2_{aY})}} \tag{8.44}$$

where Cov_a is the additive genetic covariance, and σ^2_{aX} and σ^2_{aY} are the additive genetic variances of the two traits. The additive genetic covariance is estimated in much the same manner as the additive genetic variance. For example, the additive genetic covariance of two traits is one-fourth the phenotypic covariance of half sibs.

I Genetic Correlations among Enzyme Activities

A class of quantitative traits that has aroused considerable interest among population and molecular geneticists is the level of activity of enzymes in intermediary metabolism. These enzymes are coded by well-defined structural genes, and by using balancer chromosomes and other genetic means, it is possible to examine the contributions of structural variation and background or "modifier" genes on levels of enzyme activity. In

a series of papers, Laurie and co-workers examined the genetic variability in the activities of 23 enzymes among a series of second chromosome and third chromosome replacement lines (Laurie-Ahlberg et al. 1980, 1982, 1985). (Chromosome replacement lines are genetically identical for all but one chromosome.)

The design of the experiments was fairly elaborate with each line tested on a number of days and from several different rearing bottles. The complex design was necessary in order to determine whether the differences among lines were due to genetic differences, or to rearing effects, or to measurement errors in a particular assay. Significant variability was seen among lines in the activity of all 23 enzymes in both the second and the third chromosome replacement series, with the exception of TPI (triose phosphate isomerase) in the second chromosome series. Essentially the same results were obtained whether activities were expressed as reaction rates on a per fly, per unit weight or per unit protein basis. Wilton et al. (1982) examined the patterns of correlation among the activities of pairs of the 23 enzymes, and found that there was a tendency for positive genetic correlations in enzyme activity.

A particularly striking positive correlation has been seen repeatedly in the activities of the two leading enzymes in the pentose phosphate shunt, glucose 6-phosphate dehydrogenase (G6PD) and 6-phosphogluconate dehydrogenase (6PGD). The pentose shunt directs the flux of carbon chains from glucose 6-phosphate through steps that generate NADPH, which is important in many biosynthetic steps, including lipid synthesis. Flies that lack activity of either G6PD or both of these enzymes can survive. However, flies lacking activity for 6PGD alone cannot survive. This suggests that the intermediate generated by G6PD may accumulate to lethal levels, analogous to many human inborn errors of metabolism. In this particular case the presence of both enzymes in the same pathway suggests an adaptive mechanism for coordinate control. In other cases, the positive correlations in activity may simply be related to effects on common substrates. For example, the enzymes hexokinase, G6PD, PGI and PGM, all share glucose 6-phosphate as a substrate and all show positive intercorrelations in activity. The prevalence of the correlations is seen in the common patterns found among geographically isolated populations and among the second and third chromosome lines.

Correlated response occurs when characters other than directly selected traits respond. The magnitude of correlated responses to selection is related to the additive genetic correlation between the selected and correlated characters. The expected response is expressed by the equation

$$CR_Y = ih_X h_Y r_a \sigma_{pY} \qquad \textbf{(8.45)}$$

where X is the directly selected character, with narrow-sense heritability h^2_X, i is the intensity of selection, h^2_Y is the heritability of the correlated character Y, r_a is the genetic correlation, and σ_{pY} is the phenotypic standard deviation of character Y. In words, Equation 8.45 says that, all else being equal, a doubling in the genetic correlation will double the magnitude of the correlated response. Box J presents data from a selection study that yielded an elaborate picture of correlated responses.

J Correlated Response to Selection on Weight Gain in Rats

An artificial selection experiment was carried out for 23 generations that measured weight gain in rats between the third and ninth week. (Baker et al. 1975). Six strains were characterized for 17 skull traits in order to examine patterns of correlated responses (Atchley et al. 1982). The characters included such measurements as skull length, skull width, interorbital width, braincase depth, mandible width and jaw length. The experiment consisted of two unselected control groups, two lines selected up for weight gain, and two lines selected down for weight gain, with all lines originating from a common stock. The direct response was remarkable for the magnitude of asymmetry: the males selected for greater weight gain were 3.46 standard deviations larger than the controls, while the males selected for less weight gain were 1.30 standard deviations smaller than the controls. Similarly, females selected for greater weight gain were 1.46 standard deviations larger, and females selected for less weight gain were 1.02 standard deviations smaller. Asymmetric response is generally attributed to opposing natural selection, and no more specific mechanism can be offered here. Correlated responses showed similar degrees of asymmetry. However, while the replicate lines showed statistical consistency in the magnitudes of *direct* responses, 35 of 51 comparisons between males in replicate lines were significantly different in the *correlated* response. The irreproducibility of correlated response may be due to the greater sensitivity of genetic covariances to the perturbing effects of selection and random drift.

Atchley et al. (1982) performed multivariate analyses of the correlated characters and obtained canonical variates (linear combinations of the characters that account for the greatest variance). In comparing the heritability of the individual characters to the heritability of the canonical variates, they found that the latter generally had higher narrow-sense heritability. This finding underscores the arbitrariness of the way characters are measured and gene action is inferred, since the pattern of genetic correlation suggests that the genetic variation affects suites of characters in coordinated fashion. Rather than imagining genes affecting skull length and jaw width, the data suggest that complex patterns of traits are the targets of gene action.

One final point regarding genetic correlation is that, like heritability, it is a statistical description based on phenotypic observations. An underlying genetic model is hypothesized, and support for the model depends on how well it fits the data. The best support comes from an accurate ability of the model to predict the magnitudes of correlations among various degrees of familial relationships and correlated responses to selection.

Two primary mechanisms producing genetic correlation are pleiotropy and linkage. With pleiotropy, genes that affect one character also affect others. Alternatively, the genes affecting two traits might be entirely distinct but nevertheless generate genetic correlations because of linkage disequilibrium. Since genes are aligned together on chromosomes, selection of one gene

must affect allele frequencies of nearby genes, and linkage disequilibrium can be generated. In practice, it is very difficult to distinguish the effects of pleiotropy from linkage disequilibrium, and experimental methods that attempt to disentangle these two mechanisms generally do so by attempting to control or eliminate linkage disequilibrium.

SELECTION ON MULTIVARIATE CHARACTERS

The complexities of examining evolutionary change in multiple characters were apparent in Box J, and it might seem that a complete description of multiple-trait selection would be hopelessly complex. However, the theoretical problem becomes tractable by making certain assumptions. Most important is that the multidimensional distributions of phenotypes and of genetic effects are assumed to be multivariate Gaussian. If the initial distribution is multivariate Gaussian, and selection favors the reproduction of some phenotypes over others, then the distribution would be expected to depart from Gaussian after the selection. Although this is true, if the selection is sufficiently weak, then the phenotypic distribution will remain approximately Gaussian—at least in the short run.

Multivariate Selection Theory

Although the equations for response to selection on more than one character were derived for agricultural breeding earlier (e.g. Young and Weiler 1960), Lande (1979) brought the models to the attention of evolutionary geneticists and extended the theory in many ways that make it more useful to population biologists. We will follow Lande's notation in deriving the formulation of multivariate selection response. Although the concepts are quite simple, students unfamiliar with matrix notation and matrix manipulation might find it a rough go.

First let the measurement (phenotypic value) of the ith character be z_i, and the mean of the character in the population be $\bar{z}_i$. A set of n phenotypic measures can be written as $\mathbf{z} = (z_1, z_2, z_3, \ldots, z_n)^T$, where the boldface $\mathbf{z}$ is a common notation for a vector. The superscript T indicates the "transpose" of the vector, which means that the vector is usually written as a column rather than a row of numbers. The vector of mean phenotypes is $\bar{\mathbf{z}}$. Just as the expression $P = G + E$ indicates that the phenotype is composed of genetic and environmental effects, for multivariate phenotypes the equivalent expression is

$$\mathbf{z} = \mathbf{x} + \mathbf{e} \qquad (8.46)$$

where $\mathbf{x}$ is a vector of genetic effects for the n characters, and $\mathbf{e}$ is a vector of respective environmental effects. In Box K the formulation of the multivariate Gaussian distribution is shown to be compactly expressed by the vector of means and the variance-covariance matrix. $\mathbf{P}$ is the variance–covariance matrix for the phenotypic distribution, and $\mathbf{G}$ and $\mathbf{E}$ are the variance–covariance matrices for genetic and environmental effects. We must also assume that the joint distribution of genetic and phenotypic effects is multivariate Gaussian (Bulmer 1980; Barton and Turelli 1987).

K The Multivariate Gaussian Distribution

The univariate Gaussian density function is

$$p(z) = \frac{1}{\sigma\sqrt{2\pi}} \exp[-(z-\bar{z})^2/2\sigma^2]$$

which indicates that the fraction of the population falling between a and b is

$$\int_a^b p(z)\, dz.$$

For the multivariate Gaussian density, define a variance-covariance matrix $\mathbf{P}$ as

$$\begin{pmatrix} \mathrm{Var}_1 & \mathrm{Cov}_{1,2} & \mathrm{Cov}_{1,3} & \cdots & \mathrm{Cov}_{1,n} \\ \mathrm{Cov}_{2,1} & \mathrm{Var}_2 & \mathrm{Cov}_{2,3} & \cdots & \mathrm{Cov}_{2,n} \\ \mathrm{Cov}_{3,1} & \mathrm{Cov}_{3,2} & \mathrm{Var}_3 & \cdots & \mathrm{Cov}_{3,n} \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ \mathrm{Cov}_{n,1} & \mathrm{Cov}_{n,2} & \mathrm{Cov}_{n,3} & \cdots & \mathrm{Var}_n \end{pmatrix}$$

Then the multivariate Gaussian density function can be written as

$$p(\mathbf{z}) = \frac{1}{(2\pi)^{n/2}\,|\mathbf{P}|^{1/2}} \exp[-(1/2)(\mathbf{z}-\bar{\mathbf{z}})^{\mathrm{T}}\mathbf{P}^{-1}(\mathbf{z}-\bar{\mathbf{z}})]$$

where $|\mathbf{P}|$ is the determinant of $\mathbf{P}$, and $\mathbf{P}^{-1}$ is the inverse of $\mathbf{P}$ (i.e. $\mathbf{P}^{-1}$ is such that $\mathbf{P} \times \mathbf{P}^{-1} = \mathbf{I}$, the identity matrix). Using this density function, the fraction of the population that falls between a and b of character 1, c and d of character 2, . . . , and between s and t of character n is

$$\int_a^b \int_c^d \cdots \int_s^t p(\mathbf{z}) dz_1\, dz_2 \ldots dz_n$$

The multiple integral from $-\infty$ to $+\infty$ has the value 1. Hence the multivariate Gaussian distribution is fully specified by the vector of means $\bar{\mathbf{z}}$ and the variance–covariance matrix $\mathbf{P}$.

Multiple-trait selection is modeled by assigning each phenotype a probability of survival (fitness). Note that this model is different from truncation selection, where there was a strict cutoff between being selected or not. The mean fitness is calculated as

$$\overline{w} = \int p(\mathbf{z})w(\mathbf{z})d\mathbf{z} \tag{8.47}$$

where the integration is taken over all elements of the vector $\mathbf{z}$ and $\overline{w}$ is simply the average fitness of all individuals in the population. The mean fitness expressed as a function of phenotypic means, $\overline{w}(\overline{z}_1, \overline{z}_2, \ldots, \overline{z}_n)$, is roughly analogous to Wright's adaptive topography, with one very important distinction. In Wright's adaptive topography, the axes are gene frequencies, while here the axes are the means of the phenotypic measures.

The vector of selection differentials acting between the egg and adult stage is

$$\mathbf{S} = \overline{w}^{-1} \int \mathbf{z}p(\mathbf{z})w(\mathbf{z})d\mathbf{z} - \overline{\mathbf{z}} \tag{8.48}$$

Because the environmental effects have a mean of zero, the mean phenotype after selection is

$$\overline{\mathbf{z}}' = \overline{w}^{-1} \int \mathbf{x}\, g(\mathbf{x})\, \tilde{w}(\mathbf{x})\, d\mathbf{x} \tag{8.49}$$

where $g(\mathbf{x})$ is the distribution of genetic effects, and $\tilde{w}(\mathbf{x})$ is the mean fitness of genotype $\mathbf{x}$. The change in the vector of mean phenotypes is

$$\Delta\overline{\mathbf{z}} = \overline{\mathbf{z}}' - \overline{\mathbf{z}} \tag{8.50}$$

The gradient operator, ∇ (pronounced "del") is defined as the partial derivative of each successive element of a vector, so that $\nabla F(\mathbf{x}) = (\partial F(\mathbf{x})/\partial x_1, \partial F(\mathbf{x})/\partial x_2, \ldots)$. The gradient operator yields a vector that points in the direction of steepest ascent up a multidimensional function. Applying the del operator to the mean fitness function, the fitness gradient becomes

$$\nabla \ln(\overline{w}) = \mathbf{P}^{-1}\,\mathbf{S} = \mathbf{G}^{-1}\,\Delta\overline{\mathbf{z}} \tag{8.51}$$

where $\mathbf{G}$ is the genetic variance–covariance matrix. Equation 8.51 is more intuitively clear when written as $\Delta\overline{\mathbf{z}} = \mathbf{G}\mathbf{P}^{-1}\mathbf{S}$, since it is analogous to the univarite selection response equation $R = h^2S$. $\Delta\overline{\mathbf{z}}$ is the vector of mean phenotypic changes, or the selection response, $\mathbf{S}$ is the vector of selection differentials, and $\mathbf{G}\mathbf{P}^{-1}$ is a matrix whose elements reflect the fraction of phenotypic variance and covariance that is additive genetic variance and covariance. The analogy between $\mathbf{G}\mathbf{P}^{-1}$ and h^2 is simplest when all genetic covariances are zero; then $\mathbf{G}\mathbf{P}^{-1}$ is a matrix whose diagonal is the narrow-sense heritabilities of the traits.

Multiplying Equation 8.51 through by **G**, the change in the mean vector of phenotypes under weak multivariate Gaussian selection can be written

$$\Delta \bar{\mathbf{z}} = \mathbf{G} \nabla \ln \bar{w} \qquad (8.52)$$

For example, if **G** were the identity matrix, then the mean phenotype would follow the path of steepest ascent up the mean fitness surface. Genetic covariance results in the trajectory of the population departing from the steepest ascent. The mean fitness under this model would ultimately be maximized, but the population would not get to the maximum by the fastest route if traits affecting fitness were genetically correlated.

In considering the analogy between Equations 8.52 and 8.10 for the response to selection, the same caveats apply. In the case of a single character, the response depends on the heritability and the selection differential, and as long as these remain the same, the predicted response should continue. With multiple characters, the selection differential and the additive genetic variance–covariance matrix must remain constant. Similarly, short-term response to selection often fits the prediction equation rather well, because short-term selection (the first ten generations) does not usually change the genetic variance significantly. However, the prediction equation should not be applied to selection occurring over evolutionary time scales. For the same reasons, Equation 8.52 should not be applied to evolutionary time scales, although, to illustrate a point, Lande (1976) placed bounds on the strength of selection using Equation 8.52. Evidence shows that the genetic variance-covariance matrix differs among closely related taxa (Lofsvold 1986, 1988; Kohn and Atchley 1988) and even among populations within a species (Coyne and Beecham 1987), although there are statistical problems in comparing variance–covariance matrices. Further details of the derivation of Equation 8.52 and its implications can be found in Lande (1979).

Inference of Selection from Phenotypic Data

The results of models for selection on multiple phenotypes can be used to infer the operation of natural selection, but in this application one must be especially careful of the assumptions of the models, and in addition some serious statistical problems must be overcome. Either by sampling a population at two different times, or by doing a cross-sectional study, changes in the phenotypic distribution resulting from differential mortality of the different phenotypes can be detected (Lande and Arnold 1983; Arnold and Wade 1984a,b). Changes in phenotypic distributions reflect the operation of natural selection, despite the lack of any proven changes in gene frequency. Human birth weight is easily shown to be a character that is under the influence of natural selection, because the mortality rate of very small and very big babies

is higher than the mortality of babies close to the population average (Karn and Penrose 1951). Extremes of body weight also have the highest mortality in elderly humans (Harris et al. 1988). Phenotypic selection will result in changes in gene frequency to the extent that the selected phenotypes are heritable, and selection of traits with low heritability will have minimal immediate effects on the genetic composition of the population.

Lande and Arnold (1983) show that the directional selection gradient can be estimated from the partial regression coefficients of relative fitness on each of the traits. Relative fitness itself might be estimated by the mortality suffered during the interval between two times of censusing. A classical example that illustrates the method is the data from Bumpus (1899), involving 136 sparrows that had been incapacitated by a severe winter storm. Eight measurements were taken of every bird, and about half of them subsequently revived. The phenotypic measures and the fraction of birds that survived provided a way to examine the relation between the phenotypic measures and the fitness. First, the data needed transformation to the form of a multivariate Gaussian distribution. After various confounding factors (such as age) were shown to have little effect on mortality, standard methods were used to estimate the partial regression coefficients reflecting the directional selection gradient. These coefficients were standardized to units of phenotypic standard deviation. The weight character was found to have a significant regression in males ($\beta' = -0.27 \pm 0.09$) and in females ($\beta' = -0.52 \pm 0.25$). However, the regression coefficients are negative because, contrary to expectation, the smaller birds had a lower mortality. Similar methods were applied more recently by Gibbs and Grant (1987) to a species of Darwin's finches on the Galapagos Island of Daphne Major. Generally, larger body size of birds is favored during dry years, evidently because the most abundant food consists of large, hard seeds. After the onset of El Niño, there occurred a year with ten times the normal rainfall, and a reversal in the selection differential favored small body size. The reversal was consistent with the abundant appearance of small seeds produced by adventitious plant growth.

The Bumpus method for inferring phenotypic selection has faced criticism on a number of grounds. Mitchell-Olds and Shaw (1987) pointed out that, as in other applications of multiple regression, the traits can be intercorrelated (multicolinear), the estimators may be biased when traits are measured with random experimental error, and the estimators are not consistent (do not converge to true values with increasing sample size) when the errors are not identically distributed for all traits. Lande and Arnold (1983) discussed the additional problem of a strongly selected character that is not among those studied. In the end, Mitchell-Olds and Shaw (1987) recommend experimental manipulation to accompany purely observational regression analysis of selection.

Turelli (1988) emphasized that estimates of selection differentials over long periods of time depend on the stability of the genetic variance-covariance matrix. Without direct information on the stability of **G**, it is difficult to assess the confidence in estimates of selection. Price and Grant (1985) used resampling to make many estimates of selection differentials from the same data. This generated large confidence intervals on the estimates, but did not eliminate evidence for selection.

The effect of a quantitative character on reproductive fitness can be obtained directly in some circumstances. When the component of selection is mating success, a simple comparison of phenotypes of mating and nonmating individuals constitutes such a test. For example, Taylor and Kekić (1988) measured the wing lengths of 55 male *Drosophila melanogaster* that were captured while mating and 55 nonmating males that were caught by random sweep netting. The mean wing lengths were 1.418 ± 0.012 mm and 1.372 ± 0.013 mm respectively, indicating that the mating males had significantly larger wings. Careful laboratory studies have verified the mating advantage of males with larger wings (Partridge et al. 1987; Wilkinson 1987).

Random Genetic Drift and Phenotypic Evolution

An alternative to estimating selection differentials acting on particular characters directly is to construct and test a null hypothesis. A natural null hypothesis is that the magnitude of observed changes in phenotype is no different than would be expected to occur by random genetic drift. As Turelli et al. (1988) point out, depending on the time scale, different assumptions will apply to the null hypothesis. Under very short time scales, in the realm of microevolutionary changes, it is reasonable to assume that the magnitude of genetic changes is small, so that heritabilities and genetic correlations are reasonably constant. On the other hand, when one is interested in differences associated with speciation (macroevolution), it is likely that large genetic changes have occurred which are associated with changes in the genetic variance-covariance matrix, so that the null hypothesis requires a model that predicts an equilibrium between the accumulation of new mutations and the loss of genetic variance by random genetic drift.

Considering first the macroevolutionary problem, we need to predict how the additive genetic variance may change because of sampling effects associated with finite population size. In a discrete population of effective size N_e the expectation of the additive genetic variance changes over successive discrete generations according to

$$\mathrm{E}(\sigma_a^2)' = [1 - 1/2N_e]\,\mathrm{E}(\sigma_a^2) + \sigma_m^2 \qquad \textbf{(8.53)}$$

where σ_m^2 is the increment in genetic variance added each generation by

mutation (Clayton and Robertson 1955, 1957; Lande 1979, 1980; Turelli et al. 1988). When the influx of new mutations (inflating the variance) is balanced by the loss of variance due to random genetic drift, the population arrives at what Turelli et al. (1988) call the mutation drift equilibrium. The expected genetic variance at the mutation-drift equilibrium is

$$E(\sigma_a^2) = 2N_e\sigma_m^2 \tag{8.54}$$

As in the case of random genetic drift for one gene, the population is expected to arrive at the mutation-drift equilibrium after $4N_e$ generations. Moreover, given two lineages that diverged t generations ago, the expected difference between the mean phenotypes is $2t\sigma_m^2$. Just as the neutral mutation theory predicts that the rate of gene substitution (and hence divergence) is independent of population size and depends only on the neutral mutation rate, the neutral rate of phenotypic divergence depends only on the rate of mutations affecting the phenotype. This result was derived rigorously by Lynch and Hill (1986), who also showed that the rate of divergence is dependent only on the rate of purely additive mutations, and is independent of dominance and epistatic effects. Figure 27 shows the increase in within-population variance (from an initial population with no variation), until the population reaches a steady-state balance of mutation and drift. Simulations also verify the nearly linear increase in variance between populations with time. Extending neutral divergence to the case of multiple characters, Lynch and Hill (1986) show that the neutral divergence of the variance-covariance matrix depends only on the mutational variance-covariance matrix. In a model with n genes, k alleles at each, population size N, and mutation rate μ per locus, Cockerham and Tachida (1987) found that the initial rate of increase of variance between populations depended on N, but the asymptotic rate of divergence, and the steady state variance between populations, depended only on the mutation rates.

Turelli et al. (1988) used the mutation-drift equilibrium as a null hypothesis for devising a statistical test. Evolution has been too rapid to be explained by the neutral model (with 95% confidence) if

$$\sigma_m^2/\sigma_p^2 < \frac{(z/\sigma_p)^2}{2t(1.96)^2} \tag{8.55}$$

where z is the number of phenotypic standard deviations that the mean has changed in t generations. In other words, the mutation-drift equilibrium places a lower bound on the ratio of mutational variance to the total phenotypic variance. If the observed σ_m^2/σ_p^2 is smaller than the critical value given in Equation 8.55, then mutation does not introduce enough variance to account for the observed divergence. Lynch (1988b) did a comprehensive review of experimental estimates of σ_m^2/σ_p^2 and found that it generally falls in

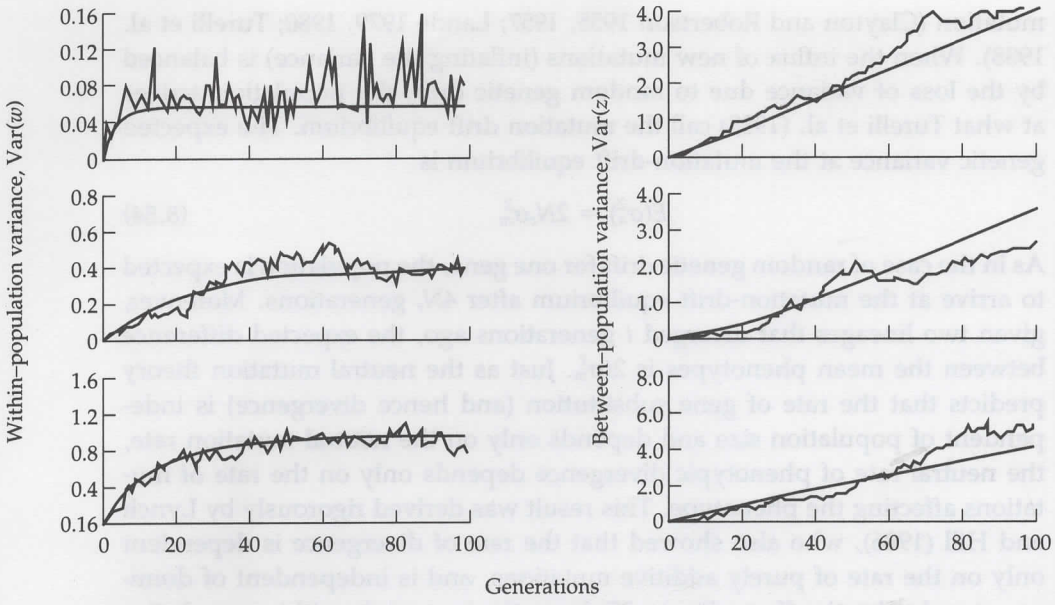


Figure 27. Simulations of a model with mutation of genes underlying a quantitative character and random genetic drift in a finite, subdivided population. The jagged lines represent the simulations, while the smooth curves are based on an analytical model. One hundred populations were run in each sample and each case. (a) Number of breeding adults, $N = 2$; number of loci, $n = 50$; mutation rate per locus, $\mu = 0.001$; additive genetic variance, $\sigma_a^2 = 0.2$; degree of dominance, $\bar{d} = 0$; dominance variance, $\sigma_d^2 = 0$; free recombination. (b) $N = 10$, $n = 10$, $\mu = 0.001$, $E(o^2) = 1$, $d = 0$, $\sigma_d^2 = 0$, free recombination. (c) $N = 10$, $n = 50$, $\mu = 0.001$, $\sigma_a^2 = 0.2$, $d = \bar{a}$, $\sigma_d^2 = 0$, ten chromosomes with five evenly spaced loci on each, recombination frequency = 0.01 for adjacent loci. Note that the variance within populations comes to a steady-state balance between mutation and loss of variation by random genetic drift, while the variance among populations continues to increase. (From Lynch and Hill 1986.)

the range of 10^{-2} to 10^{-4}. The rate of mutation is adequate to maintain variability in life history parameters in the cyclic parthenogen *Daphnia* (Lynch 1985).

The test in Equation 8.55 is only useful on a macroevolutionary time scale, when there is some hope that the diverged populations actually have reached a mutation-drift equilibrium. For situations like Bumpus's sparrows or the influence of the rains caused by El Niño on Darwin's finches, this test is not appropriate. In these cases, the approach that Lande (1976, 1977) devised is more reasonable, although it too must be applied cautiously. Lande assumed that heritability was constant, which is plausible for microevolutionary changes, but as we shall see, it presents an immediate contradiction. Under the assumption of constant heritability, the expected mean phenotype re-

mains constant but the variance increases as $t\sigma_a^2/N_e$. Thus, constant heritability forces the rate of phenotypic drift to vary inversely as the effective population size. As Turelli et al. (1988) point out, this is in contradiction to the neutral mutation theory, which, as with the mutation-drift model of phenotypic change, predicts that the rate of neutral divergence depends on mutation rates and not population size. Nevertheless, because the constant heritability model is more appropriate for microevolutionary tests, the contradiction is not necessarily fatal. Lande (1976) noted that drift can be rejected at the 5% level as a cause of phenotypic divergence if

$$N_e > \frac{(1.96)^2 h^2 t}{(z^2/\sigma_p^2)} \tag{8.56}$$

If the effective population size is larger than this critical value, then the divergence is too great to be explained by drift. An example provided by Lande (1976) concerned phenotypic changes in the teeth of two species of fossil horses. Although the mutation-drift equilibrium model is more appropriate for this analysis, Turelli et al. (1988) found that neither test rejected the null hypothesis, which implies that the observed rate of phenotypic evolution was not more rapid than could be explained by neutral drift. The point is not that morphological variation is neutral (there is a wealth of evidence in Chapter 9 to the contrary), but rather that the rates of morphological evolution are generally so slow that, by themselves, neutrality cannot be rejected.

EVOLUTION OF QUANTITATIVE TRAITS: MUTATION-SELECTION BALANCE

Stabilizing selection is of great interest for the role it may play in the persistence of heritable genetic variation in multifactorial traits. Intuitively, it seems reasonable to suppose that observed levels of additive genetic variation may result from a balance between stabilizing selection, which tends to reduce genetic variation, and new mutations, which tend to increase it. Deceptively simple when stated verbally, the models become complex when reduced to mathematics, and include such complications as the number of genes, the type of action of alleles and their interactions, linkage between genes, the type and intensity of selection, and the influence of selection on other traits which are related through pleiotropy. Pleiotropy is reflected in the widespread tendency of genes to affect several traits simultaneously, sometimes including seemingly unrelated traits, which usually results from the fact that complex phenotypic traits are determined by the interactions of the products of many genes during development. The number of genes

affecting a trait is relevant to mutation–selection balance because, as Crow and Kimura (1970) showed, for a given genetic variance, the selection intensity per locus decreases as the number of loci increases. If the total mutation rate is fixed, the per-locus mutation rate must decrease as the number of loci increases. For a population in mutation-selection balance, the intensity of selection per locus is approximately independent of the number of loci (Turelli 1984). Complex as the situation is, the hypothesis that additive genetic variation might be maintained by stabilizing selection balanced against mutation has been supported by mathematical approximations in the models studied by Lande (1975, 1977, 1979, 1980).

On the other hand, the approximations appear to be quite sensitive to the particular genetic and mathematical assumptions that are made. The simplest models ignore dominance and epistasis, and they assume a large number of alleles per gene, the distribution of effects of which are approximated by a normal or Gaussian distribution. Assuming a Gaussian distribution of allelic effects at individual loci is a very different thing from assuming a Gaussian distribution of phenotypes, as the latter is open to empirical validation.

With these caveats in mind, let us examine the models that predict a balance between mutational increase in variance offset against a loss of variance by optimizing selection. Following Kimura (1965), Lande (1975), and Turelli (1984), let $p_t(x)$ be the distribution of allelic effects before selection in generation t. In each generation, selection occurs, then mutation, and then reproduction. Assume that phenotypes and underlying genotypic effects have a Gaussian distribution; since selection would be likely to change the distribution, this approximation remains acceptable only under weak selection. Although the model as developed is haploid, it is intended as an approximation for a diploid model assuming additivity. Letting $p_t'(x)$ be the density of allelic effects after selection, μ be the mutation rate (the fraction of gametes that are mutated), and $g(x)$ be the density of mutational effects on the phenotype, the recursion for $p(x)$ is

$$p_{t+1}(x) = (1 - \mu)p_t'(x) + \mu \int_{-\infty}^{+\infty} p_t'(y)g(x - y)dy \qquad \textbf{(8.57)}$$

This model, called the "Kimura-Lande-Fleming" model by Turelli (1984), partitions the phenotypic distribution into two components. A fraction $(1 - \mu)$ of the distribution derives from nonmutant genes, and their distribution is $p_t'(x)$ after selection. Fitness follows a Gaussian distribution with an optimum phenotype coinciding with the mean. Genes that mutate change phenotype x to y with distribution $g(x - y)$, and then undergo selection, resulting in the distribution $p_t'(y)$. Since any phenotype x can mutate to many different values of y, the integral is taken over all values of y. Kimura (1965) showed that, with weak selection, the process in Equation 8.57 comes to a

steady state with a Gaussian distribution of allelic effects. Lande (1975) stated the condition for an approximate Gaussian equilibrium as being $m^2 << \sigma_g^2$, where m^2 is roughly the variance in mutational effects, and σ_g^2 is the genetic variance.

On the other hand, Turelli (1984) showed that, in order to achieve the equilibrium genetic variance predicted by the Kimura-Lande-Fleming model, since the magnitude of the effects of mutations must be restricted in order to make the Gaussian approximation valid, then the assumed *rate* of mutation must be unacceptably high (above 10^{-4} per gene per generation). An alternative approximation, based on a model first presented by Kingman (1978), avoids the necessity of assuming that mutational effects are weak, by assuming that allelic effects after mutation are independent of allelic effects before mutation. Turelli (1984) performed numerical simulations of the Kimura-Lande-Fleming model, and found that the Gaussian approximation gave better predictions of the equilibrium variance under conditions of weak mutational variance, high mutation rates and weak selection; while Turelli's "house-of-cards" approximation gave more accurate predictions with lower mutation rates, higher mutational variance, and stronger selection.

Extensions of models that incorporate multivariate selection and mutation have also been considered (Lande 1980; Turelli 1985, 1986). Fitnesses are assumed to have a multivariate Gaussian distribution, implying optimizing selection, and mutational effects are also assumed to be multivariate Gaussian. Alternative assumptions are found to be invalidated when selection also occurs indirectly, through correlated responses on traits related by pleiotropy (Turelli 1986). Pleiotropy can even obscure the relation between the intensity of selection and the magnitude of the additive genetic variance. When pleiotropy occurs, an increase in the intensity of selection may decrease the additive genetic variance, or it may increase it, depending on the particular assumptions that are invoked—for example, whether stabilizing selection occurs on each of the pleiotropically related traits, or whether directional selection in antagonistic directions favors alternative sets of alleles.

The conclusion seems to follow that the simple mathematical models require so many artificial assumptions that predictions based on them cannot inspire great confidence (Turelli 1986). This difficulty is inherent in evaluating whether observed values of additive genetic variance and heritability can be explained by mutation-selection balance. All the more does it complicate the wider goal of attempting evolutionary explanations of the relative amounts of the variance in phenotype which are contributed by additive gene effects, dominance, interactions between genes, and environmental components.

The genetic architecture also includes genetic and environmental correlations among different traits, which result from pleiotropy and common environmental effects. Genetic variation can be maintained by selection acting

alone, when the favorable effects of alleles in respect to some traits are offset by unfavorable pleiotropic effects on different traits. Negative pleiotropic effects that result in offsetting selection pressures are known as **antagonistic pleiotropy**. An important implication of antagonistic pleiotropy is that the action of natural selection cannot maximize every trait simultaneously, because favorable changes in some traits bring about unfavorable changes in others, and adaptive responses in some directions are nonadaptive in others. When natural selection acts on traits that are negatively correlated, the best that can usually be achieved is a compromise, a sort of best choice of bad bargains. For example, in *Drosophila*, there is antagonistic pleiotropy in females between the mean rate of egg laying and longevity (Rose and Charlesworth 1981a), and selection for increased female longevity decreases the mean rate of egg laying (Rose and Charlesworth 1981b).

Hastings (1987) cast the question of mutation-selection balance in a different light. He observed that the phenomenon of selection appears ubiquitous at the phenotypic level, but at the molecular level, neutrality fits reasonably well. This may prove to be the single most important enigma in modern population genetics: that evolution seems to follow different rules at these two levels. Theoreticians are at work to reconcile the paradox, and empirical work, focused on identifying genetic markers of quantitative traits, has already met with some success. Hastings (1987) specifically addressed stabilizing selection on a quantitative character as it affects the rate of substitution of alleles of genes that determine the trait. Simulations were performed with 8 and 16 genes, each with two alleles contributing equally and additively to the trait. For biologically reasonable parameters, the rates of substitution were found to be as low as 1/10 the neutral rate (with $4N_e s = 4$). On the other hand, it would be very difficult to infer selection based on substitution rate alone, since the rate of substitution depends strongly on the number of genes affecting the quantitative trait (Foley 1987).

MENDELIAN MARKERS AND QUANTITATIVE TRAITS

Despite the landmark paper of Fisher (1918) and the impressive work since, there remains a gap between the theory of quantitative traits and the identification of classical Mendelian genes. The good fit of molecular variation to many aspects of the neutral theory contrasts strongly with the appeal of adaptation in morphological evolution. Although the discrepancy is no doubt exaggerated by intuitive notions that something as obvious as horns must be useful, and something as subtle as silent nucleotide substitutions must have negligible effects, it remains an important problem to bring these two levels of population genetics together. One approach is to identify the explicit

relations among Mendelian genes and quantitative traits. The Mendelian genes might be identified by molecular function, but this is not essential. The two methods discussed below focus on mapping genes affecting quantitative traits and on the response of particular genes to selection of continuous characters.

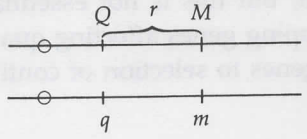
Mapping Genes that Influence Quantitative Characters

There has been a renewed interest in mapping genes (or blocks of genes) that affect quantitative traits largely because of the density of markers identified by restriction fragment length polymorphisms. These recombinant DNA methods might be of direct impact in plant and animal breeding programs, as well as in the diagnosis and study of human disease. Observations of linkage between genes that influence quantitative traits and major genes date back to Sax (1923). Interest in actually mapping "quantitative trait loci" (or QTL's) also has a long history, beginning with Thoday's (1961) studies of bristle traits in *Drosophila*. Reviews of the early work can be found in Thoday (1979). The essence of the experiments on *Drosophila* was to construct lines that differed with respect to a quantitative character (such as up and down selected lines for bristle number), and to construct a random set of recombinants among them, using multiply marked chromosomes (bearing several recessive mutations whose phenotype can be readily scored), to infer which chromosomal regions the recombinants carried. The end result of such experiments is not a genetic map in the classical sense, identifying locations of a series of specific genes that determine the quantitative character. The result is rather a statistical description of the character that indicates what fraction of the genetic variance in the parental lines is contributed by each region identified by the flanking markers.

The simplest experimental design for mapping QTL's is shown in Figure 28. Two lines that differ in one or more quantitative traits are crossed and the F_1 are subsequently crossed to produce many F_2 progeny. The Mendelian markers can be identified by any means including discrete, segregating traits, allozyme markers, or restriction site polymorphisms. The best chance of success is expected with larger numbers of traits and larger numbers of markers. In the F_2 of parental lines that were homozygous for alternative marker alleles of one gene, the genotypes are expected to occur in a 1:2:1 ratio. Means and variances of these three genotypes are calculated and compared using linear models. Figure 28 shows the expected phenotypes of the three genotypes in the F_2 generation, and shows not only that the means should differ, but that the variances should differ as well.

The simple logic of this approach has considerable appeal, and it is remarkable that it often works in practice. For example, using this design in

(a) Hypothetical F_1 genotype **Generates F_2 array**



F_1 gametes	Frequencies
QM	$\frac{1}{2}(1-r)$
qm	$\frac{1}{2}(1-r)$
Qm	$\frac{1}{2}r$
qM	$\frac{1}{2}r$

Genotypes	Frequency	Value
$QQMM$	$\frac{1}{4}(1-r)^2$	$+a$
$QqMM$	$\frac{1}{2}(r-r^2)$	d
$qqMM$	$\frac{1}{4}r^2$	$-a$
$QQMm$	$\frac{1}{2}(r-r^2)$	$+a$
$QqMm$	$\frac{1}{2}[r^2+(1-r)^2]$	d
$qqMm$	$\frac{1}{2}(r-r^2)$	$-a$
$QQmm$	$\frac{1}{4}r^2$	$+a$
$Qqmm$	$\frac{1}{2}(r-r^2)$	d
$qqmm$	$\frac{1}{4}(1-r)^2$	$-a$

(b) Marker-locus class means (frequency-adjusted)

Marker class	Mean expression
MM	$(1-2r)a + 2r(1-r)d$
Mm	$[(1-r)^2 + r^2]d$
mm	$(1-2r)(-a) + 2r(1-r)d$

(c) Expressions to resolve additive and dominance effects

Additivity: $(MM - mm)/2 = a(1-2r)$
Dominance: $Mm - (MM + mm)/2 = d(1-2r)^2$

Figure 28. A design for interpreting the relationship between marker locus genotypes and their phenotypes through the effects of a linked locus that affects the quantitative trait. (a) The F_1 genotype is heterozygous for the quantitative trait locus Q and the marker M, with a frequency of recombination of r. Gametes of the $F_1 \times F_1$ cross are combined to give the nine two-locus genotypes, their respective frequencies among the F_2, and their genotypic values. The mean phenotype for each of the three marker genotypes can be expressed as functions of additive (a) and dominance (d) effects at a quantitative trait locus (b). These allow estimation of additive and dominance effects based on the phenotypes of the marker genotypes (c). (From Edwards 1987.)

crosses between two species of *Lycopersicon* (tomato), Weller et al. (1988) examined 18 phenotypic attributes and 10 markers. Of the 180 possible marker-trait combinations, 85 (47%) showed significant effects on the mean, and 75 (42%) were significant at the 0.1 to 1% level. (Since many tests are done simultaneously, one expects to get "significant" results by chance alone. One common way to minimize this problem is to make the cutoff level of significance more restrictive.) In an even larger study, Edwards et al. (1988) examined 82 traits and 17–20 segregating marker genes in the F_2 of two lines of maize. They found that most traits showed a significant influence with about two-thirds of the markers, and that, taken together, the marker genes could explain between 8% and 40% of the phenotypic variance of the traits examined in detail. Each marker explained from 0.3% to 16% of the phenotypic variance. Digenic analysis was also done to see whether markers and

the genetic regions they identified resulted in epistatic interactions in the quantitative traits. Although dominance was not uncommon, epistatic interactions were rare. These experiments provide a unique view of components of genetic variance, but it must be realized that only a small sample of genotypes are used in constructing the F_2 lines, so they cannot accurately reflect components of variance in the large panmictic population from which they were drawn. Similar methods have been applied with comparable success by Burr et al. (1988), Tanksley et al. (1982) and by Kluge and Geldermann (1982). More details on the theory behind these methods are given in Jayakar (1970), Beckmann and Soller (1983, 1986), Weller (1986) and George and Elston (1987).

Although the method of scoring F_2 individuals is powerful, it is not essential to perform controlled crosses to make inferences about the effects of marker genes on quantitative characters. This is fortunate because methods for identifying marker genes relevant to polygenic human diseases would be of great utility. The chances for success in identifying QTL's in humans is much improved by selecting Mendelian markers that may have a functional relation to the trait (*candidate genes*). For example, Sing and Davignon (1985) examined a sample of humans for apolipoprotein E (*apoE*) genotype and several quantitative traits relating to serum triglyceride, cholesterol, and low-density lipoproteins. There are three common alleles of *apoE*, yielding six genotypes, the mean phenotypes of which are presented in Table 10. About 16% of the total genetic variance in blood serum levels of LDL-C (an important carrier of cholesterol) is explained by genotype at the *apoE* locus. The importance of this kind of approach is underscored by the observation that about half of the variation among individuals in serum cholesterol is associated with polygenic variation.

Selection for Continuous Traits and Response of Identified Genes

During the height of the neutralist-selectionist controversy in the 1970s, many population geneticists tried to find evidence of direct selective effects of allozyme variants. This problem was essentially the same as in the preceding section, except that the focus was on finding *the* target of selection rather than merely linked markers. This restricted focus resulted in the needs for all kinds of genetic gymnastics to randomize the genetic background and to impose functionally and mechanistically relevant selection on the target gene-enzyme system. One promising approach used selective media, analogous to selective media used to isolate bacterial mutants. Three examples of this approach in *Drosophila* include selection of amylase allozymes using starch medium (DeJong and Scharloo 1976; Hickey 1979; Clark and Doane 1984), selection of G6PD allozymes using octanoate in the medium (Bijlsma 1978),

Table 10. Serum lipid levels in a sample of 102 people from Ottawa, adjusted for age, sex, height, and weight effects.

	GENOTYPE						Probability (*F* ratio)	ESTIMATES OF THE POPULATION	
	ϵ4ϵ4	ϵ3ϵ3	ϵ2ϵ2	ϵ4ϵ3	ϵ3ϵ2	ϵ4ϵ2		Mean	Variance
Count	4	63	2	21	10	2	—	—	—
Relative frequency	0.039	0.618	0.020	0.206	0.098	0.020	—	—	—
Variable									
Total cholesterol	180.3	173.8	136.0	183.5	161.4	178.1	0.09	174.16	732.48
HDL-C	53.3	47.3	47.1	47.3	45.7	45.2	0.93	47.32	130.43
LDL-C	102.9	104.2	73.9	112.8	89.5	109.3	0.08	104.00	602.17
VLDL-C	24.0	22.3	15.0	23.3	26.2	23.5	0.83	22.83	127.56
Triglycerides	65.5	74.4	70.6	70.4	79.3	73.5	0.96	73.60	918.01
LDL-Apo B	86.3	83.9	55.8	86.8	78.4	60.5	0.13	83.03	375.72
VLDL-Apo B	8.5	11.4	5.65	10.5	9.2	18.9	0.56	10.91	66.21
LDL-C/total cholesterol	0.56	0.60	0.55	0.61	0.55	0.61	0.26	0.59	0.0055
VLDL-C/total triglycerides	0.37	0.32	0.19	0.33	0.33	0.32	0.79	0.32	0.0200

(From Sing and Davignon 1985.)

and selection of alcohol dehydrogenase allozymes using ethanol in the medium (Cavener and Clegg 1981a, 1981b). In the last example, two cage populations with normal medium, and two cage populations with 10% ethanol medium, were maintained for 57 generations. The cages were started with progeny from 11 matings of wild-caught flies. Samples were taken periodically and ten enzyme polymorphisms were monitored. Since sizes of the samples ranged from 90 to 120 genomes, the allele frequency estimates were fairly reliable. Figure 29 shows the frequency of the Adh^s allele during the selection, revealing a pronounced response favoring the Adh^f allele in the ethanol medium. This result is consistent with the greater activity of the Adh^f flies. In addition to the *Adh* response, allozymes of GPDH, MDH and 6PGD also showed significant response to the ethanol selection.

Using a device called an "inebriometer" (illustrated in Figure 30), Cohan and Graf (1985) tested *Drosophila* from nine populations on the west coast of the United States and found that the more northerly populations could endure ethanol fumes for a longer time before passing out from ethanol intoxication. The time that it takes a fly to fall to the bottom of the series of

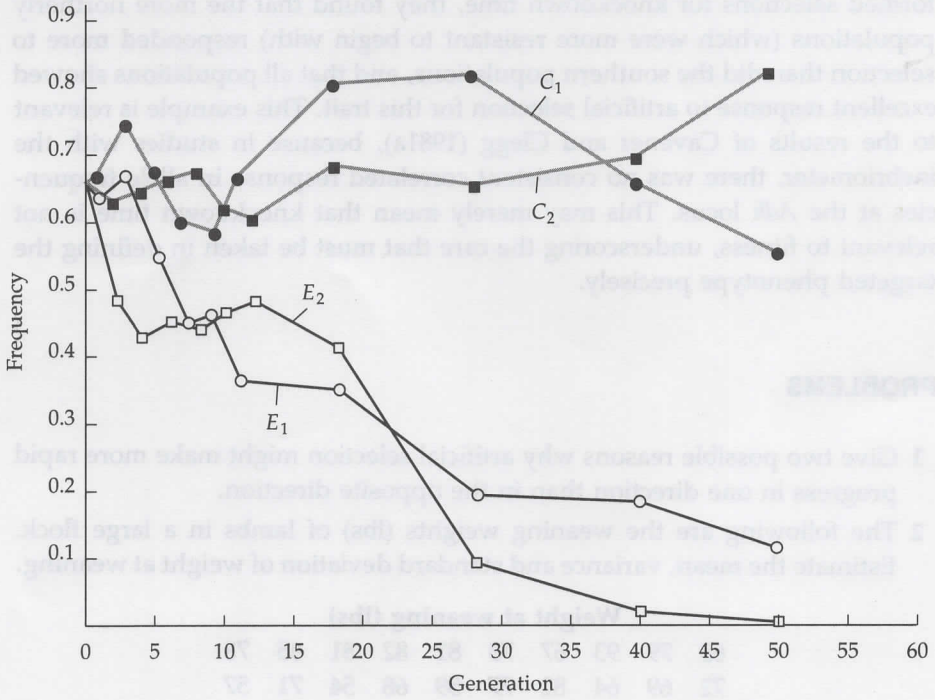


Figure 29. Adh^S frequencies for two control replicate populations (C_1 and C_2) and two ethanol-selected replicate populations (E_1 and E_2). (From Cavener and Clegg 1981a.)

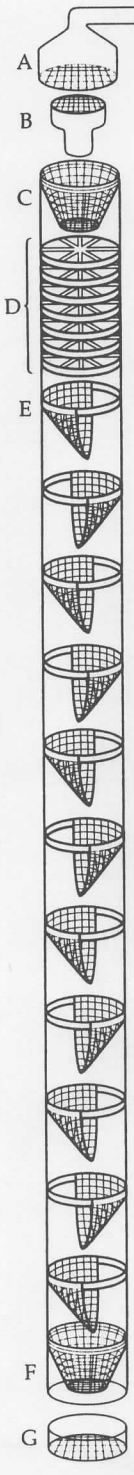


Figure 30. An inebriometer, a device for measuring knockdown resistance to ethanol fumes. A, Removable top of the knockdown tube, connected to a flow of ethanol fumes. B, Cartridge for inserting flies into the tube. C, Funnel for holding the cartridge in place. D, Diffuser for separating the mass of flies. E, Baffles to slow down the fall of the flies. The baffles ensure that flies that have only partially succumbed to the fumes have many chances to regain their footing. F, Funnel for concentrating flies on the collecting dish. G, The collecting dish onto which flies tumble when they can no longer keep their grip on the baffles. (After Cohan and Graf 1985.)

baffles is the "knockdown time," and flies with a greater tolerance of ethanol fumes have greater knockdown times. When Cohan and Graf (1985) performed selections for knockdown time, they found that the more northerly populations (which were more resistant to begin with) responded more to selection than did the southern populations, and that all populations showed excellent response to artificial selection for this trait. This example is relevant to the results of Cavener and Clegg (1981a), because in studies with the inebriometer, there was no consistent correlated response in allele frequencies at the *Adh* locus. This may merely mean that knockdown time is not relevant to fitness, underscoring the care that must be taken in defining the targeted phenotype precisely.

PROBLEMS

1 Give two possible reasons why artificial selection might make more rapid progress in one direction than in the opposite direction.

2 The following are the weaning weights (lbs) of lambs in a large flock. Estimate the mean, variance and standard deviation of weight at weaning.

Weight at weaning (lbs)									
68	79	93	67	73	81	82	81	85	78
72	69	64	82	77	59	68	54	71	57
88	97	69	60	92	62	64	64	90	60

3 Suppose a population of *Drosophila* has a normal distribution of abdominal bristles with a mean of 20 and a standard deviation of 2. What proportion

of the population is expected to fall into the following categories of bristle number:

a. Between 18 and 22?
b. Greater than 22?
c. Greater than 24?
d. Between 20 and 22?
e. Smaller than 16?

4 Two inbred varieties of tobacco are crossed and give a variance in leaf number in the F_1 generation of 1.5. The variance in the F_2 generation is 6.0. What are the genotypic and environmental variance components and the broad-sense heritability in the F_2 population?

5 In a population of the flour beetle *Tribolium* the mean weight of pupae is 2000 μg. The phenotypic variance is 40,000 μg^2, and the additive genetic variance is 10,000 μg^2. If individuals with a mean pupal weight that is 2 phenotypic standard deviations above the mean are selected, what is the expected average pupal weight among the progeny?

6 If a population of *Drosophila* has a mean number of abdominal bristles of 20 with a narrow sense heritability of 25 percent, what is the expected bristle number after one generation when the selection differential is 4 bristles? What is the expected number after 10 generations of equally intense selection?

7 Five generations of selection for decreased plasma cholesterol level in mice decreased the mean from 2.16 mg/100 ml to 2.01 mg/100 ml. The average selection differential was 0.07 mg/100 ml. What is the realized heritability?

8 A quantitative trait has the following mean values for the genotypes shown:

AA	AA'	$A'A'$
23.8	25.2	19.4

a. What are the values of a and d?
b. What allele frequency of A would maximize the mean value of the trait in a panmictic population?

9 Two strains of rats were selected for increased or decreased pigmentation on the head and back. After 10 generations the high strain had a rating of 3.73 and the low strain a rating of −2.01. The strains were crossed, and the standard deviations in the F_1 and F_2 generations were 0.87 and 0.60, respectively. Estimate the effective number of factors affecting the trait in these strains.

10 Estimate the correlation coefficient in litter size between first and second litters using the following data from 10 females.

First litter	8	9	9	10	10	10	11	11	13	13
Second litter	6	8	12	10	12	12	9	10	12	12

11 How many generations of selection with a selection differential of 20 would be required to increase the average number of eggs laid per hen per year from 180 to 220, given a heritability of 20%?

12 A standard normal distribution is a normal distribution with mean 0 and variance 1. Using this distribution, calculate the mean of the top 1/2, 1/4, 1/8, 1/16, and 1/32 of the population. These values are the standardized selection intensities (i) for the proportions saved, which correspond to truncation points of 0.00, 0.68, 1.16, 1.54, and 1.86, respectively.

13 If the selection differential differs in males and females, show that the proper value to use in Equation 8.10 is the mean.

14 Consider a locus with genotypes AA, AA', and $A'A'$ whose contribution to a quantitative trait has $a = 0.6$ and $d = 0.2$; another locus with genotypes BB, BB', and $B'B'$ contributes to the same trait with $a = 0.4$ and $d = 0$. If the loci are unlinked and additive and the allele frequencies of A and B are 0.5, and 0.7, respectively, calculate the narrow-sense and broad-sense heritability of the trait when the total phenotypic variance is 1.0.

15 A herd of dairy cattle yields milk with a fat content of 3.4% ± 0.65% (mean ± standard deviation) and a protein content of 3.3% ± 0.45%. The heritabilities of these traits are 60% and 70%, respectively, and the genetic correlation is 0.55. If selection is practiced for percent protein with a selection intensity of $i = 1.5$, what increase in percent protein and percent fat would be expected? What intensity of selection would produce the same increase in percent fat by direct selection?

16 Show that a simple Mendelian dominant gene has a narrow sense heritability of $2(1 - q)/(2 - q)$, where q is the frequency of the dominant allele.

17 For an overdominant locus with two alleles, show that the additive genetic variance at equilibrium equals 0.

18 The intensity of selection i is the mean of the selected parents in a standard normal distribution when B is the proportion saved. It equals the selection differential in units of standard deviation, so $i = (\mu_S - \mu)/\sigma$. Over the range $B = 0.05$ to $B = 0.005$, i is given approximately by $i = 0.8 + 0.41 \times \ln [(1/B) - 1]$ (Simmonds 1977). Calculate i for $B = 1/2$, 1/4, 1/8, 1/16, and 1/32; compare with the corresponding values in Problem 12.

19 If a normally distributed phenotype X, with truncation point T, is transformed to a standard normal $x = (X - \mu)/\sigma$, with transformed truncation

point $t = (T - \mu)/\sigma$, what is the relationship between the ordinate Z of the untransformed distribution at T to the ordinate z of the standardized distribution at the point t?

20 Define $N(x) = [1/\sqrt{(2\pi)}]\exp(-x^2/2)$ as the normal distribution of a phenotype with mean 0 and variance 1. Use the fact that $dN(x)/dx = xN(x)$ to show that the mean μ_S of all phenotypes greater than t satisfies $\mu_S = z/B$, where $z = N(t)$ and B is the proportion of the population with phenotypes greater than t.

Chapter Nine

Ecological Genetics and Speciation

The population genetic models that have been considered up to this point are remarkable in the simplicity of their assumptions. The models leave out many variables that could affect gene frequencies in actual cases. Reducing the genome to a one-gene, two-allele model is a mind-boggling simplification from the point of view of information theory, since some genomes may contain billions of bits of information (three billion nucleotides in the case of humans), and the models focus on only one pair of bits. Still, the models have been useful in distilling the essence of the effects of evolutionary forces.

Perhaps the least reductionist of the biological sciences is ecology. Ecology aims to understand the factors that determine the distribution and abundance of populations and communities of organisms. So many factors may be operative, from rainfall to temperature to chemical interactions among species, that it is a real accomplishment to prove any one factor is critical. The four principal forces in population genetics — mutation, migration, natural selection, and random genetic drift — are all more complicated when ecological variables intrude. For example, mutation rates depend on dietary intake of chemical mutagens and exposure to radiation sources. Migration rates depend on local population densities and the availability of food. Natural selection is most intimately tied to ecology, since the growth and regulation of population number is often analogous to the concept of fitness as used in population genetics. Indeed, as you will see, the Malthusian parameter (which expresses the intrinsic rate of population increase in ecological models) can be equated with the conventional Wrightian fitness used in population genetics. Random genetic drift also has an ecological component in that the magnitude of the effects of random genetic drift depends on popu-

lation number. The importance of random genetic drift in small local demes was emphasized in Wright's shifting balance theory, and when random drift yielded genotypes of superior fitness, natural selection took over and the demes containing the favored gene combinations expanded in size. Therefore, the ecological component of regulation of population size was actually implicit to the shifting balance theory.

Overall rates of morphological evolution are closely tied to ecology. Judged on the basis of the fossil record, observed rates of evolution vary enormously. Different groups of organisms evolve at very different rates; a single group of organisms can evolve more rapidly at one time and less rapidly at another; and different traits can evolve at different rates even within a single group of organisms over a given period. Such variation in evolutionary rates implies that the limiting factor is not the actual mode of evolution — whether evolution proceeds mainly by the shifting balance process, the mass selection process, or whatever. Rather, the rate of evolution is limited by the interaction between ecological opportunity, in the form of special environmental conditions, and genetic variation having the potential to respond successfully to the ecological opportunity (Stebbins and Hartl 1988).

Most species are usually constrained by their physical environment and by other organisms. (In this context, the word "constraint" means "limitation.") Physical constraints include, for example, seasonal weather; the vast majority of plants and animals find winter inhospitable and have evolved behavioral or physiological adaptations to survive the cold periods. Actual physical barriers may also be physical constraints; for fish and water lilies, for example, the shoreline of a lake is such a barrier. Physical constraints also include availability of essential nutrients in soil or water.

Biological constraints are those associated with the presence of other organisms. For example, predation of the young is a limiting factor in fish population growth; excess shade caused by a forest canopy limits sunlight available for germination and growth of green plants in the understory; and infectious disease often decimates populations that have become too dense.

The evolution of most species, most of the time, merely keeps pace with their ever-changing physical and biological constraints. Such offsetting evolutionary changes are often called "treadmill evolution;" more colorfully this situation is sometimes called the "Red Queen" hypothesis after the Queen of Hearts in Lewis Carroll's *Through the Looking Glass*, who tells Alice, "Now, here, you see, it takes all the running you can do, to keep in the same place" (Van Valen 1973). Note that a constraint must be changing — intensifying — for treadmill evolution to occur. A steady-state constraint may eventually be overcome by adaptations resulting from natural selection, provided that the necessary genetic variation becomes available through mutation.

Occasionally an opportunity for rapid evolution becomes available. A geographical constraint may disappear or be surmounted, resulting in the colonization of a new, favorable habitat such as an island. Or a biological constraint may disappear, as when a major predator, competitor, or parasite becomes extinct. Or, in less striking situations, the rate of intensification of a constraint may become insufficient to offset natural selection. Reduction of constraints partly releases a species from the treadmill, and it becomes free to evolve so as to expand its ecological range. In extreme cases, evolution can be very rapid and result in a spectacular adaptive radiation of many newly formed species. One exceptionally well-studied case — the explosive radiation of *Drosophila* species in the Hawaiian islands — resulted in the evolution of at least 400 endemic species in the several million years since the formation of the archipelago by volcanic action (see Carson et al. 1970; Richardson 1974; Carson and Kaneshiro 1976; Templeton 1979; Beverly and Wilson 1985). Thus, most species appear to have ample reserves of genetic variation with the potential to enable rapid evolution in new directions, but their rate of evolution is limited by the absence of ecological opportunity (Stebbins and Hartl 1988).

Ecological considerations have been incorporated into population genetics theory in a number of ways, some of which are illustrated in this chapter. Perhaps the most direct way to connect ecology and population genetics is through models of density-dependent selection, in which one set of equations determines population regulation, another set determines gene frequency change, and the equations are interdependent. One exciting outcome of this theory is that it generates hypotheses that are amenable to experimental test.

A second connection between ecology and population genetics is through population regulation. Populations do not often occur in large panmictic units, but show various levels of subdivision. The ecological component of population subdivision is whether population number is regulated at the level of the local deme or the whole population. The level at which population regulation occurs makes the distinction between "hard" and "soft" selection, discussed later in this chapter. Ecology and population genetics also share an interest in demography in attempting to understand population growth in organisms characterized by age-specific rates of reproduction and mortality. Incorporating demographic theory into population genetics has been technically demanding, but some elegant principles have emerged, and predictions from theories about the evolution of senescence are currently being tested.

During the 1970s, when protein electrophoresis was in its heyday, many investigators sought physiological mechanisms for the maintenance of enzyme polymorphisms. Since the postulated mechanisms often involved specific adaptations to the environment, the experimental studies often had an

ecological component. Two examples of questions approached in this manner were: (1) Is aminopeptidase-1, which shows a marked cline in allele frequencies in mussel populations of Long Island Sound, involved in osmoregulation? (2) Do different forms of globins in field mice, which happen to be correlated with the elevation above sea level at which the populations live, result from adaptations to thin air? There are risks in assuming that every feature of an organism exists for some adaptive purpose — an assumption that has been called the Panglossian paradigm (Gould and Lewontin 1979). The risk is in attributing adaptive functions where none exist, a logical fallacy aptly summarized by Mark Twain, who observed that our legs have evolved to be just long enough to reach the ground. Despite the philosophical problems arising from the strong intuitive appeal of adaptive explanations (Gould and Lewontin 1979), attempts to find physiological and ecological reasons for polymorphisms have had the beneficial effect of unifying aspects of population genetics and ecology.

DENSITY-DEPENDENT NATURAL SELECTION

The concept of fitness used up to this point is different from that used in ecological models. In population genetics, the definition of fitness is the rate of transmission of one's genes, relative to that of other genotypes in the population. A healthy, fertile genotype might have a fitness of 0.5 instead of 1 only because another genotype is reproducing twice as fast. In demographic theory, the parameter most closely related to fitness is the Malthusian parameter, which determines the absolute rate of increase in size of the population. Demographic theory was originally developed with the implicit assumption that the population was genetically homogeneous (no genetic variation), with all individuals having the same rates of reproduction and death. Although the Malthusian parameter is an absolute measure of fitness, ratios of absolute fitnesses can be used for most purposes in the same way as relative fitnesses.

Models bringing together population dynamics and population genetics were not considered until 1971 (Roughgarden 1971; Charlesworth 1971; Anderson 1971). Traditionally, changes in population size were thought to occur in **ecological time**, that is, over the course of a few generations. Changes in gene frequencies (equated with evolutionary change) had been considered as much slower processes that occurred in **evolutionary time**. The distinction between the ecological and evolutionary time scales broke down with the observations that third chromosome inversions in *Drosophila pseudoobscura* sometimes change in frequency in annual cycles (Dobzhansky 1943). Rapid evolution of heavily pigmented forms of the moth *Biston betularia*, referred to as industrial melanism, provided additional evidence of the blurred distinction between ecological and evolutionary time scales.

Beginning with the familiar one-gene two-allele viability model, let alleles A and a have frequencies p and $q = 1 - p$, and suppose that genotypes AA, Aa, and aa have fitnesses w_{11}, w_{12}, and w_{22}. The fitnesses are the probabilities of surviving from zygote to the adult stage. The recursion for allele frequencies was derived in Chapter 4 as

$$p' = \frac{p^2 w_{11} + pqw_{12}}{\overline{w}} \tag{9.1}$$

where $\overline{w}$ is the average fitness of the population, obtained by summing the frequencies of each genotype weighted by its fitness:

$$\overline{w} = p^2 w_{11} + 2pqw_{12} + q^2 w_{22} \tag{9.2}$$

The implication of density-dependent selection is that the fitnesses w_{11}, w_{12}, and w_{22} are no longer constants but depend on the total population size. If population growth obeys the simple formulation

$$N' = \overline{w}N \tag{9.3}$$

where N is the population size, then density dependence implies that as the population size increases, say to the point of reducing the food supply, then $\overline{w}$ must decrease. This means that $\overline{w}$ is a function of population size. However, the density dependence might be different for each genotype, resulting in different genotypes having different responses to crowding. For example, one genotype might do well under crowded conditions, another might be relatively insensitive to population size, and a third might show a sharp reduction in fitness as the population increases in size. When the absolute density-dependent fitnesses are compared with one another, the relative fitnesses of the three genotypes might differ in their ranking depending on density. An equation commonly used to describe the density dependence of fitness is known in ecological theory as the **logistic equation**.

$$w_{ij} = 1 + r_{ij}(K_{ij} - N)/K_{ij} \tag{9.4}$$

The logistic equation has two parameters for each genotype A_iA_j: r_{ij}, which describes the intrinsic rate of increase (indicating how rapidly the population grows when at very low density); and K_{ij}, which is the **carrying capacity**. K_{ij} is called the carrying capacity because when the population size is $N = K_{ij}$, the second term in Equation 9.4 is zero, so $w_{ij} = 1$. If A_iA_j were the only genotype present, and $w_{ij} = 1$, then Equation 9.3 implies that $N' = N$. Hence, the carrying capacity is an equilibrium population size. To illustrate the parameters r_{ij} and K_{ij}, Figure 1 shows the relation between w_{ij} and population size for three genotypes. The subscripts on r and K specify the genotype, so each of the three genotypes has its own r_{ij} and K_{ij}. Figure 2 shows how a population containing only one genotype would grow, and the population size approaches the carrying capacity K_{ij} asymptotically.

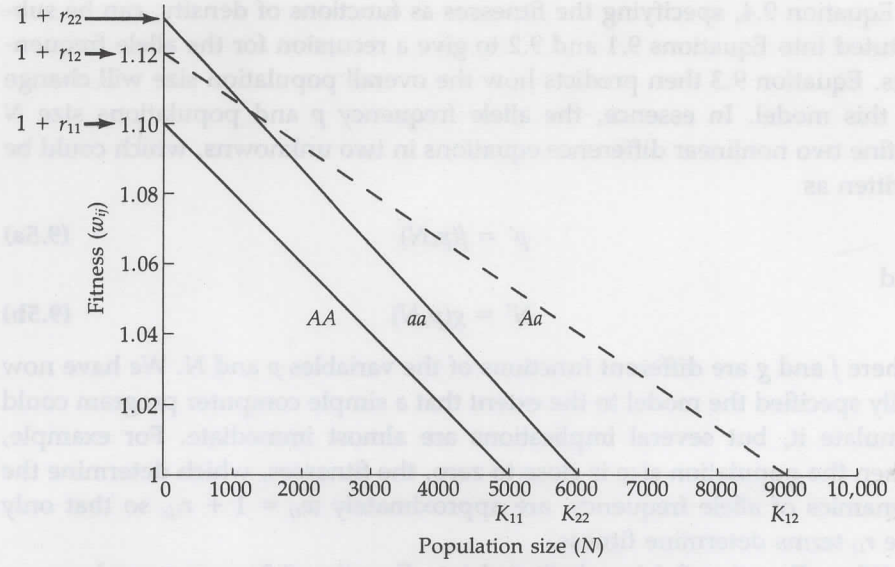


Figure 1. The relationship between fitness and population size in a logistic model of density-dependent selection. Each genotype has an associated intrinsic rate of increase (r_{ij}) and carrying capacity (K_{ij}). The parameters are $r_{11} = 0.1$, $r_{12} = 0.12$, $r_{22} = 0.13$, $K_{11} = 5000$, $K_{12} = 9000$, and $K_{22} = 6000$.

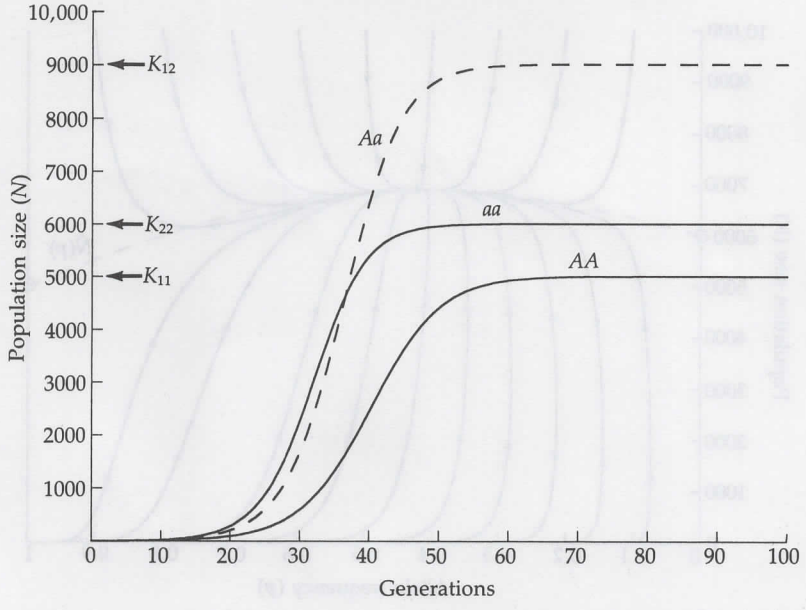


Figure 2. The dynamics of population growth for the three genotypes presented in Figure 1, using the equation $N' = \overline{w}N$. Each curve represents the growth of the population in which only the single genotype is found. Note that the rate of initial increase depends on r, and the final population size is the carrying capacity.

Equation 9.4, specifying the fitnesses as functions of density, can be substituted into Equations 9.1 and 9.2 to give a recursion for the allele frequencies. Equation 9.3 then predicts how the overall population size will change in this model. In essence, the allele frequency p and populations size N define two nonlinear difference equations in two unknowns, which could be written as

$$p' = f(p,N) \tag{9.5a}$$

and

$$N' = g(p,N) \tag{9.5b}$$

where f and g are different functions of the variables p and N. We have now fully specified the model to the extent that a simple computer program could simulate it, but several implications are almost immediate. For example, when the population size is close to zero, the fitnesses, which determine the dynamics of allele frequency, are approximately $w_{ij} = 1 + r_{ij}$, so that only the r_{ij} terms determine fitness.

When Equation 9.4 is substituted into Equation 9.3, a very cumbersome expression for $\overline{w}$ is obtained, but after some algebra, it simplifies to

$$\overline{w} = 1 + \overline{r} - (\overline{r/K})N \tag{9.6}$$

Thus the equation for the change in population size can be expressed in

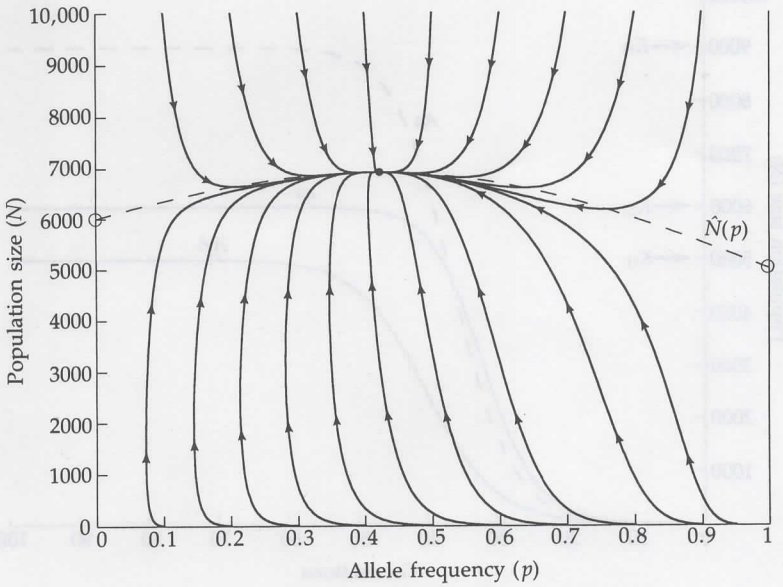


Figure 3. The $\hat{N}(p)$ curve (dashed) for the overdominant carrying capacity case given in Figures 1 and 2. Solid lines represent population trajectories, showing both the changes in population size and gene frequency. The open circles represent unstable equilibria; the filled circle represents the stable equilibrium. (After Roughgarden 1979.)

terms of the averages of r_{ij} and of (r_{ij}/K_{ij}). Following Roughgarden (1971), if p were fixed, then the population would evolve to an ecological equilibrium with $\hat{N} = \overline{w}\hat{N}$, implying that $\overline{w} = 1$. However, with $\overline{w} = 1$, Equation 9.6 implies that

$$\hat{N} = \bar{r}/(\overline{r/K}) \tag{9.7}$$

Since the $\hat{N}$ assumes a fixed allele frequency, Roughgarden (1971) labeled it $\hat{N}(p)$, and he showed that any equilibrium of the system must be either a maximum or a minimum of $\hat{N}(p)$. For a single population, any stable equilibrium must be a maximum of $\hat{N}(p)$. Using the definitions of the averages $\bar{r}$ and $\overline{r/K}$, Equation 9.7 can be rewritten

$$\hat{N}(p) = \frac{p^2r_{11} + 2pqr_{12} + q^2r_{22}}{p^2(r_{11}/K_{11}) + 2pq(r_{12}/K_{12}) + q^2(r_{22}/K_{22})} \tag{9.8}$$

To find maxima of $\hat{N}(p)$, set $d\hat{N}(p)/dp = 0$ and solve for p. The solution is algebraically cumbersome, but the analysis is straightforward. The equilibrium $\hat{p}$, which depends on the r_{ij} and K_{ij} terms, is stable if there is overdominance in carrying capacities (if $K_{11} < K_{12} > K_{22}$) and unstable if the carrying capacities show underdominance. Figures 3 and 4 show simulations of these

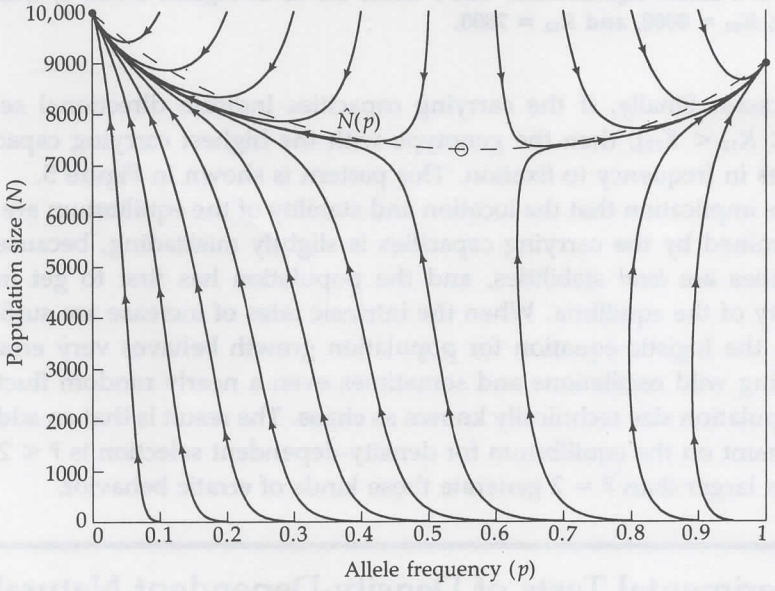


Figure 4. When the carrying capacities have an underdominant pattern, the two fixations are stable, and the population follows a course that depends on the initial population size and gene frequency. As in Figure 3, the dashed line is the $\hat{N}(p)$ curve and solid lines are population trajectories. The open circle is an unstable equilibrium and the two filled circles are stable equilibria. The r terms are as in Figures 1 and 2, but here $K_{11} = 9000$, $K_{12} = 6000$, and $K_{22} = 10{,}000$.

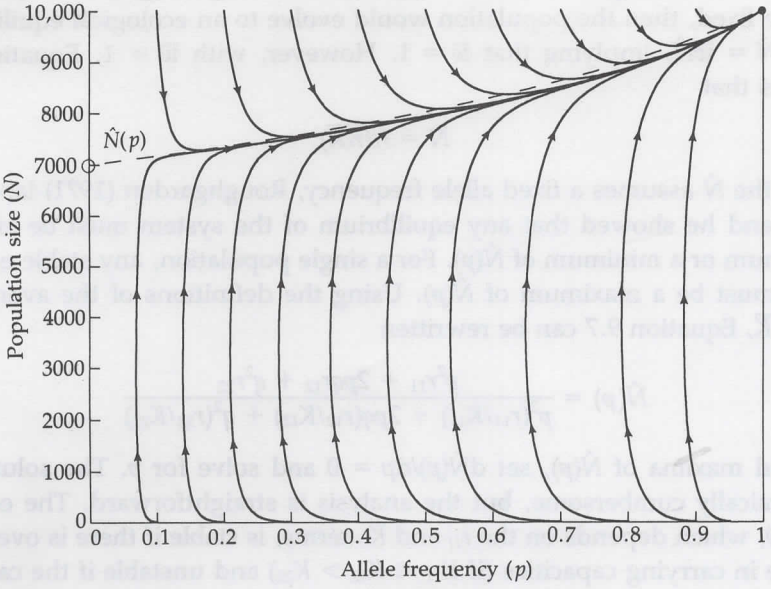


Figure 5. Density-dependent selection when one homozygote has the highest carrying capacity. As in Figures 3 and 4, the dashed line is the $\hat{N}(p)$ curve and solid lines are population trajectories. The open circle is an unstable equilibrium and the filled circle is a stable equilibrium. The r terms are as in Figures 1 and 2, with K_{11} = 10,000, K_{12} = 8000, and K_{22} = 7000.

two cases. Finally, if the carrying capacities indicate directional selection ($K_{11} < K_{12} < K_{22}$), then the genotype with the highest carrying capacity increases in frequency to fixation. This pattern is shown in Figure 5.

The implication that the location and stability of the equilibrium are totally determined by the carrying capacities is slightly misleading, because these stabilities are *local* stabilities, and the population has first to get into the vicinity of the equilibria. When the intrinsic rates of increase are sufficiently large, the logistic equation for population growth behaves very erratically, showing wild oscillations and sometimes even a nearly random fluctuation in population size technically known as **chaos**. The result is that an additional constraint on the equilibrium for density-dependent selection is $\bar{r} < 2$, since values larger than $\bar{r} = 2$ generate these kinds of erratic behavior.

A Experimental Tests of Density-Dependent Natural Selection

A fundamental issue of density-dependent selection is the trade-off between *r*-selection and *K*-selection; that is, whether a genotype can do well either at low density or at high density, but not at both. Mueller and Ayala (1981) tested for such trade-offs by estimating

intrinsic rates of increase of a set of 25 lines of *Drosophila melanogaster*, homozygous for different second chromosomes and reared at different densities. All 25 lines were reared at densities of 10, 20, 50, 100, 250, 500, 750, and 1000 adults per culture bottle, and after one week of egg laying the surviving adults were counted and discarded. From these counts, the rate of increase of each line at each density was estimated. Trade-offs in growth rate between low and high density were assessed through the correlation coefficients between the growth rates at low and high density. Perhaps surprisingly, the correlations were all positive, implying that lines that had high growth rates at low density also had high growth rates at high density. This finding contradicts the assumption that there must be a trade-off between the rates. In these experiments some lines were healthy and always reproduced at high rates, while other lines were generally sick and never reproduced well.

In a second experiment to measure the effects of natural selection on the parameters of density-dependent regulation, Mueller and Ayala (1981) reared mixed populations at a density of 100 flies per bottle (low density) or near the carrying capacity (high density). Three replicate populations at each density were maintained for five months, at which time the rates of population growth were tested at densities of 10, 500, and 750 flies per culture bottle. The results, shown in Figure 6, show that the response was in the direction expected of a trade-off, since the lines selected at low density had higher rates

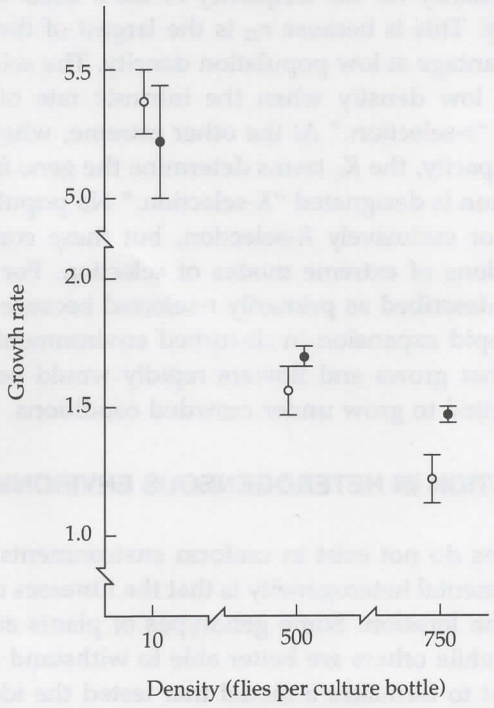


Figure 6. Mean per capita growth rate (and 95% confidence intervals) at three densities in *r*-selected (open circles) and *K*-selected (solid circles) populations of *Drosophila melanogaster*. (From Mueller and Ayala 1982.)

of growth at low density and lower rates of growth at high density. In any case, genetic variation in the populations was capable of responding to density-dependent selection. In analogous experiments, Marks (1982) examined larval survival, development rate, and female weight in *Drosophila* as functions of density, and generally found positive correlations among the components. On the other hand, many studies have shown that crowding larvae results in females with lower fecundity (e.g. Robertson 1957). Motivated by these findings, Prout and McChesney (1985) examined theoretical models with density dependence of female fecundity, and found highly complex population dynamics when both larvae and adults were censused. Clark and Feldman (1981b) emphasized the importance of density dependence on mating pairs, and experimental results showed that responses to crowding differed among mating pairs of genotypes. Because the behavior of natural populations can be complex, the correspondence to theory can be discouragingly poor. The primary result of the experimental tests of density dependence is that the intuition that there should be trade-offs between success at high and low densities was ill-founded.

When population size is kept low by means other than density-dependent growth (for example, by severe storms or drought), the r_{ij} terms may play a role in gene frequency dynamics. For example, in Figures 3, 4, and 5, there is a consistent tendency for the frequency of the a allele to increase at low population density. This is because r_{22} is the largest of the r_{ij} terms, giving the a allele an advantage at low population density. The selection that occurs in populations at low density when the intrinsic rate of increase is at a premium is called "r-selection." At the other extreme, when population size is near carrying capacity, the K_{ij} terms determine the gene frequency dynamics, and the selection is designated "K-selection." No population faces exclusively r-selection or exclusively K-selection, but these concepts have been useful as descriptions of extreme modes of selection. For example, weedy plants have been described as primarily r-selected because their life cycle is geared towards rapid expansion in disturbed environments; in a disturbed habitat, a weed that grows and flowers rapidly would be favored initially over one better suited to grow under crowded conditions.

NATURAL SELECTION IN HETEROGENEOUS ENVIRONMENTS

Natural populations do not exist in uniform environments, and one consequence of environmental heterogeneity is that the fitnesses of genotypes may differ depending on location. Some genotypes of plants are better adapted to drier habitats, while others are better able to withstand flooding. Levene (1953) was the first to formalize a model that tested the idea that heterogeneous environments might promote polymorphism. The Levene model assumes random mating among the entire population, with juveniles dispers-

ing into local demes, where they face different intensities and directions of selection. The juveniles remain in the same deme until they mature, when the adults unite once again into one panmictic unit.

Levene assumed that the number of adults in each deme was constant; this implies a form of natural selection called **soft selection** (Wallace 1968). An example of soft selection is when a large number of seeds lands in each deme, but only a fraction of the seeds can survive to maturity. For example, the plants may be trees, limited by the need for exposure to sunlight, with only a certain number of adult trees supportable in each deme. Natural selection through differential mortality of trees results in some genotypes surviving and others not, but whenever there is a death, room is made for another juvenile to mature. Hence, natural selection can occur and change gene frequencies within each deme, yet the number of adults in each deme remains constant.

The soft selection model was contrasted with **hard selection** (Dempster 1955), in which the number of zygotes landing in each deme is constant, but natural selection through mortality results in different numbers of adults. An example of hard selection might occur when seeds are distributed sparsely throughout a harsh habitat. Although some seeds survive to adulthood, ill-adapted seeds fail, and variable numbers of adults survive the selection. Implicit in the model of soft selection is competition among individuals in the same deme, which is lacking in the model of hard selection. The models of hard and soft selection are illustrated in Figure 7.

To formulate a model of hard selection, assume there are k local demes or niches, with the fitnesses of genotypes AA, Aa, and aa in niche h given by $w_{11,h}$, $w_{12,h}$, and $w_{22,h}$. Suppose that the fraction of zygotes that land in niche number h is a_h. Then the mean fitness of genotype ij is the weighted average across all niches,

$$w_{ij} = \sum_{h=1}^{k} a_h w_{ij,h} \tag{9.9}$$

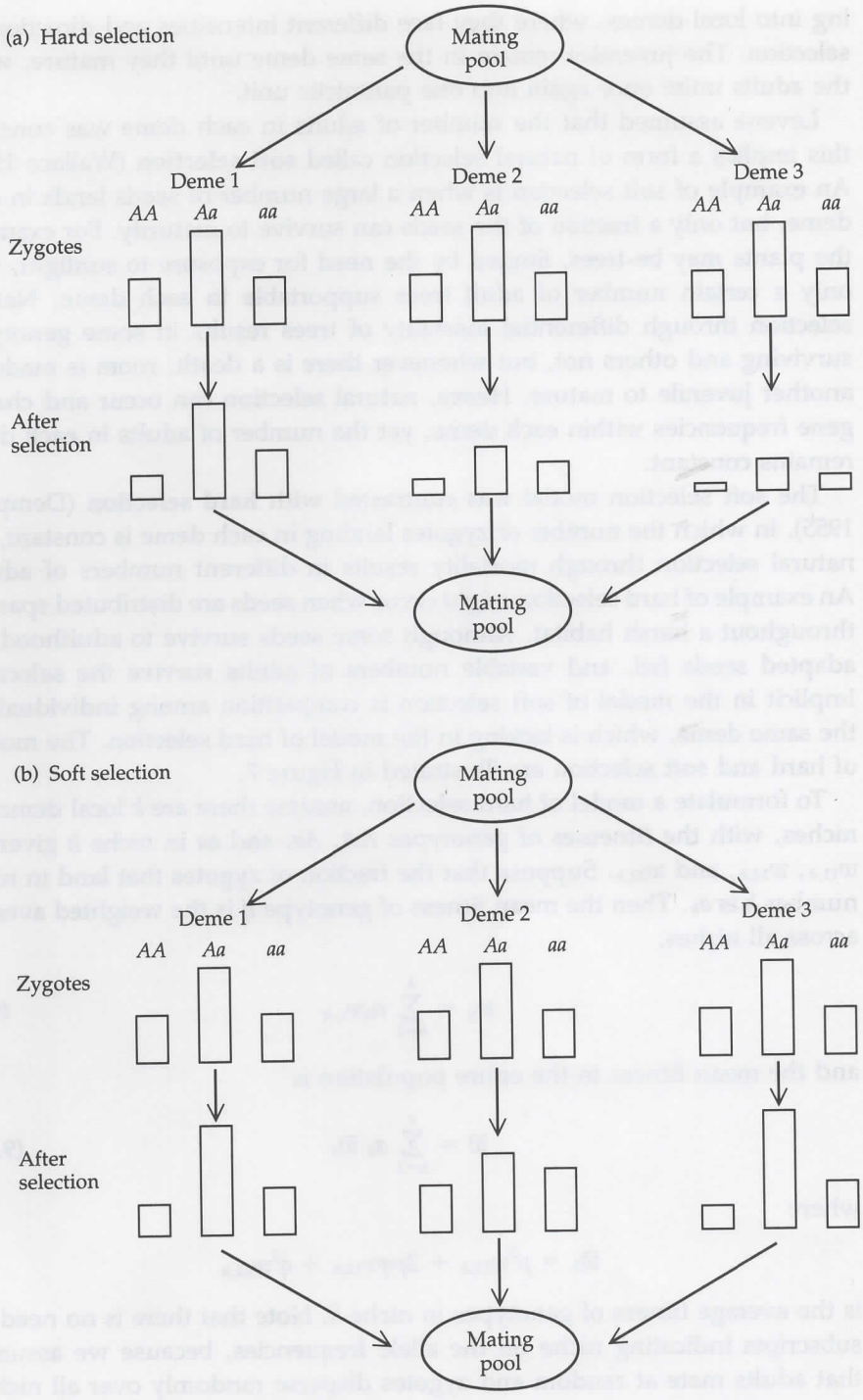
and the mean fitness in the entire population is

$$\overline{w} = \sum_{h=1}^{k} a_h \overline{w}_h \tag{9.10}$$

where

$$\overline{w}_h = p^2 w_{11,h} + 2pq w_{12,h} + q^2 w_{22,h}$$

is the average fitness of genotypes in niche h. Note that there is no need for subscripts indicating niche on the allele frequencies, because we assumed that adults mate at random and zygotes disperse randomly over all niches. Dempster (1955) showed that the change in allele frequency is


(a) Hard selection
Mating pool
Deme 1
Deme 2
Deme 3
AA Aa aa
AA Aa aa
AA Aa aa
Zygotes
After selection
Mating pool
(b) Soft selection
Mating pool
Deme 1
Deme 2
Deme 3
AA Aa aa
AA Aa aa
AA Aa aa
Zygotes
After selection
Mating pool

◁ **Figure 7. Models of hard and soft selection. The height of each bar represents the *number* of individuals of each genotype at each stage. In hard selection (a), the population randomly mates and then disperses to separate demes, where natural selection changes both the relative frequencies of genotypes and the relative sizes of demes. In soft selection (b), the population randomly mates and disperses, but natural selection does not alter the number of adults that are present within each deme after selection.**

$$p' = \frac{p^2\overline{w}_{11} + pq\overline{w}_{12}}{\overline{w}} \tag{9.11}$$

This is the familiar equation for viability selection, but the viabilities are replaced with averages taken over niches. A necessary and sufficient condition for stable polymorphism is

$$\Sigma a_h w_{11,h} < \Sigma a_h w_{12,h} > \Sigma a_h w_{22,h} \tag{9.12}$$

which is to say $\overline{w}_{11} < \overline{w}_{12} > \overline{w}_{22}$. Since classical overdominance in terms of average fitnesses across all niches is necessary for polymorphism, hard selection does not provide any increased likelihood for maintaining polymorphism. This is an important and counterintuitive result: environmental heterogeneity alone does not guarantee increased likelihood of polymorphism. Only certain forms of selection in heterogeneous environments support greater polymorphism, as shown below.

In the model of soft selection (often called the Levene model), the number of adults in each deme is constant, and demes that favor rare genotypes will increase their representation in the total population. This results in a sort of frequency dependence that does make it easier to maintain polymorphisms through habitat variability. Each generation every niche receives zygotes in Hardy-Weinberg proportions, and natural selection occurs, so that after selection in niche h

$$p'_h = \frac{p^2 w_{11,h} + pq w_{12,h}}{\overline{w}_h} \tag{9.13}$$

The change in allele frequency in the total population is the average of the p'_h, or

$$p' = \sum_{h=1}^{k} a_h p'_h \tag{9.14}$$

Equation 9.14 proves to be quite difficult to solve in general, but the conditions for a protected polymorphism are straightforward. If $dp'/dp > 1$ for both fixations, then a polymorphism is protected because the fixations would be unstable. Differentiating Equation 9.14,

$$dp'/dp = \sum_{h=1}^{k} a_h \, (dp'_h/dp) \tag{9.15}$$

where

$$\left.\frac{dp'_h}{dp}\right|_{p=0} = \frac{w_{12,h}}{w_{22,h}}$$

and

$$\left.\frac{dp'_h}{dp}\right|_{p=1} = \frac{w_{12,h}}{w_{11,h}}$$

Combining these results, the fixations are both unstable if

$$\sum_{h=1}^{k} a_h(w_{12,h}/w_{22,h}) > 1 \tag{9.16a}$$

and

$$\sum_{h=1}^{k} a_h(w_{12,h}/w_{11,h}) > 1 \tag{9.16b}$$

Since the fitnesses can be normalized to $w_{12,h} = 1$, Equations 9.16 can be written as

$$\frac{1}{\Sigma a_h(1/w_{22,h})} < 1 \tag{9.17a}$$

and

$$\frac{1}{\Sigma a_h(1/w_{11,h})} < 1 \tag{9.17b}$$

Thus polymorphism is protected under the Levene soft selection model if the harmonic mean fitness of each homozygote is less than the harmonic mean fitness of the heterozygote. Since the arithmetic mean is always greater than the harmonic mean, it is easier to obtain a polymorphism with soft selection than with hard selection. This agrees with the intuitive notion that soft selection results in a sort of frequency dependence favoring rare genotypes in the niches for which they are adapted.

Although soft selection makes it easier to obtain polymorphisms, there is relatively little chance of polymorphism in the absence of overdominance, even in the Levene model (Hoekstra et al. 1985). The ability of the Levene model to maintain polymorphism is increased with spatial subdivision of the mating population (Karlin 1982), or with nonrandom assortment of genotypes into niches. A model of nonrandom assortment, called *habitat selection* (Maynard Smith 1962), can maintain polymorphisms even with heterozygote in-

feriority (Wilson and Turelli 1986). For additional refinements of models of selection in heterogeneous environments, see Christiansen (1974, 1975), Prout (1968), and Gillespie (1974, 1975). In haploids, the conditions for maintaining a polymorphism with soft selection are the same as those for a diploid locus with dominance (Gliddon and Strobeck 1975), and the number of niches must exceed the number of alleles at equilibrium (Strobeck 1979b). Wade (1985) derived the covariance between fitness and genotype frequency, and he showed that hard and soft selection are distinguished by different partitionings of selection into within-deme and between-deme components. This approach unifies hard and soft selection with kin and group selection, discussed later in this chapter.

Temporal Variation in Selection Coefficients

B

The idea that polymorphism can be maintained by selection coefficients varying in time has also been explored. The classical formulation of one gene with two alleles was studied by Haldane and Jayakar (1963). They assumed that the fitnesses of genotypes *AA*, *Aa*, and *aa* were $w_{11,t}:1:w_{22,t}$, where t represents generation t. They were able to show that any polymorphism is protected if the geometric mean fitnesses of the two homozygotes are both less than one, regardless of the temporal pattern of fitnesses. That is, for a protected polymorphism,

$$(\prod w_{11,t})^{(1/n)} < 1 > (\prod w_{22,t})^{(1/n)}$$

where the product $\prod$ is taken over n consecutive generations. This requirement is more restrictive than the condition for protection of polymorphism with spatially heterogeneous environments because the geometric mean is always larger than the harmonic mean.

A different problem related to temporal variation in selection was solved by Kimura (1954) using a diffusion approximation (see also Gillespie 1973). Kimura considered a haploid population with genotypes *A* and *a* fluctuating in selective advantage. When averaged over many generations, the mean fitnesses were equal, so the alleles were on average selectively neutral. The population was assumed infinite in size, so in the absence of selection the gene frequencies would not change. Kimura found that fluctuating selection tends to disperse populations away from intermediate allele frequencies, so that as time progresses the distribution of allele frequencies becomes U-shaped. However, no population can ever become fixed because the populations are infinite in size.

The theory of temporally fluctuating selection has been extended in several directions, including the incorporation of mutation (Takahata and Kimura 1979; Takahata 1981) and autocorrelated selection coefficients (Hartl and Cook 1973). One important result is that the expected distribution of gene frequencies under certain patterns of fluctuating selection and mutation can mimic the mutation–neutral drift distribution (Tier 1981; Gillespie 1978).

Demonstrating the significance of fluctuating selection in the real world has proven

to be difficult. Many investigators have estimated allele frequencies in several populations sampled sequentially over several generations. Changes in allele frequency resulting from random genetic drift depend on the effective population size, and these sampling effects occur regardless of the strength of selection. With fluctuating selection, the mean allele frequencies may or may not change, but the magnitude of the oscillations in allele frequency would be greater than expected under random drift. The comparison of oscillations in allele frequency is the essence of the test that Mueller et al. (1985a,b) applied to *Drosophila pseudoobscura* and the butterfly *Euphydryas editha*. Using estimates of the effective population size, computer simulations were used to generate expected variance–covariance matrices of allele frequencies. Observed data were tested against the theoretical expectations using a χ^2 statistic, and in both cases significant departures from pure random drift were detected. However, this test assumes a closed population, which could easily be violated. Lynch (1987) applied a similar test to populations of *Daphnia*, which live in ponds and are far more likely to represent closed populations. Although the hypothesis of zero means of selection coefficients could not be rejected in the *Daphnia* populations, the allele frequency fluctuations were much greater than expected from random drift, suggesting fluctuating selection. In *Daphnia*, heterogeneity in gene frequencies among ponds is more likely caused by fluctuating selection than random drift or diversifying selection. Yet another approach to temporally varying environments is that of *F* statistics, which were used by Barker et al. (1986) to show significant fluctuations in gene frequencies in *Drosophila*.

AGE-STRUCTURED POPULATIONS AND THE EVOLUTION OF SENESCENCE

Demographic Theory

For organisms that survive and reproduce over a number of years, the theory of population growth can become rather involved. **Demography** is the branch of biology concerned with age-specific distributions of survival and reproduction, and demographic considerations are necessary when trying to estimate fitness of long-lived organisms. For example, in estimating the relative fitness of humans with a late-onset genetic disorder such as Huntington disease, nondemographic concepts — egg-to-adult viability, for example, or a single measure of fecundity — may be too simplistic. A critically important factor is the timing of the onset of the disease relative to the reproductive period.

Changes in population size result from mortality or reproduction, which are described by age-specific distributions of survival and fecundity. Figure 8 shows the distribution of survivorship l_x, defined as the probability of surviving to the middle of age class x. This curve is known as the **survivorship curve**. When mortality is low at young ages, the survivorship curve

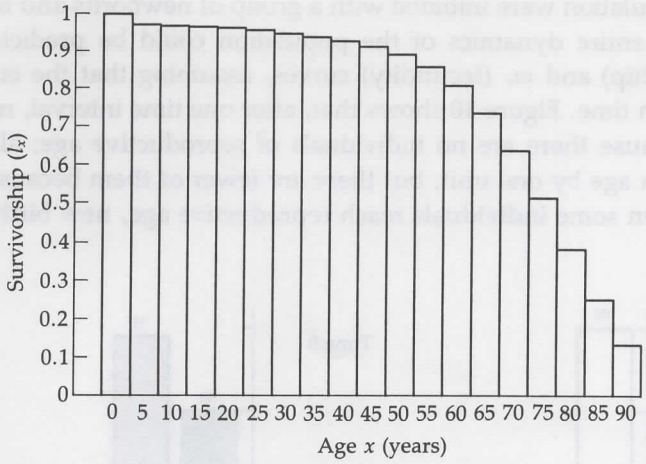


Figure 8. The survivorship curve of the human population in the United States in 1964. The estimation of these l_x values is an involved process, and there are several different ways to obtain estimates. (After Keyfitz 1977.)

remains approximately flat, until an age when mortality increases and the survivorship curve begins to drop. Increasing rate of mortality at greater age is known as **senescence**.

The demographic representation of reproduction is in terms of a **fecundity curve**, in which m_x is defined as the number of female offspring produced per female in age class x. A human fecundity curve is illustrated in Figure 9. Fecundity is zero in the prereproductive period, then increases during reproductive output, and declines again to zero in the period of postreproduction.

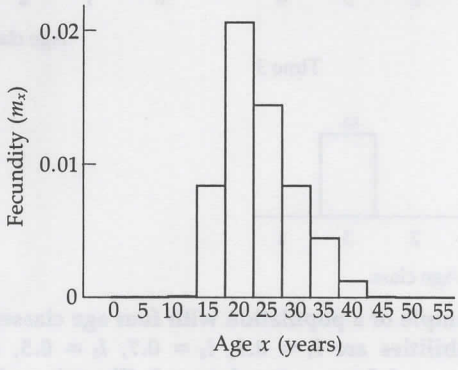


Figure 9. The fecundity curve of females in the United States in 1964. The value m_x is the average number of female births per year per female in age class x. (After Keyfitz 1977.)

If a population were initiated with a group of newborns and followed over time, the entire dynamics of the population could be predicted by the l_x (survivorship) and m_x (fecundity) curves, assuming that the curves remain constant in time. Figure 10 shows that, after one time interval, no new births occur because there are no individuals of reproductive age; all individuals increase in age by one unit, but there are fewer of them because some have died. When some individuals reach reproductive age, new births occur, and

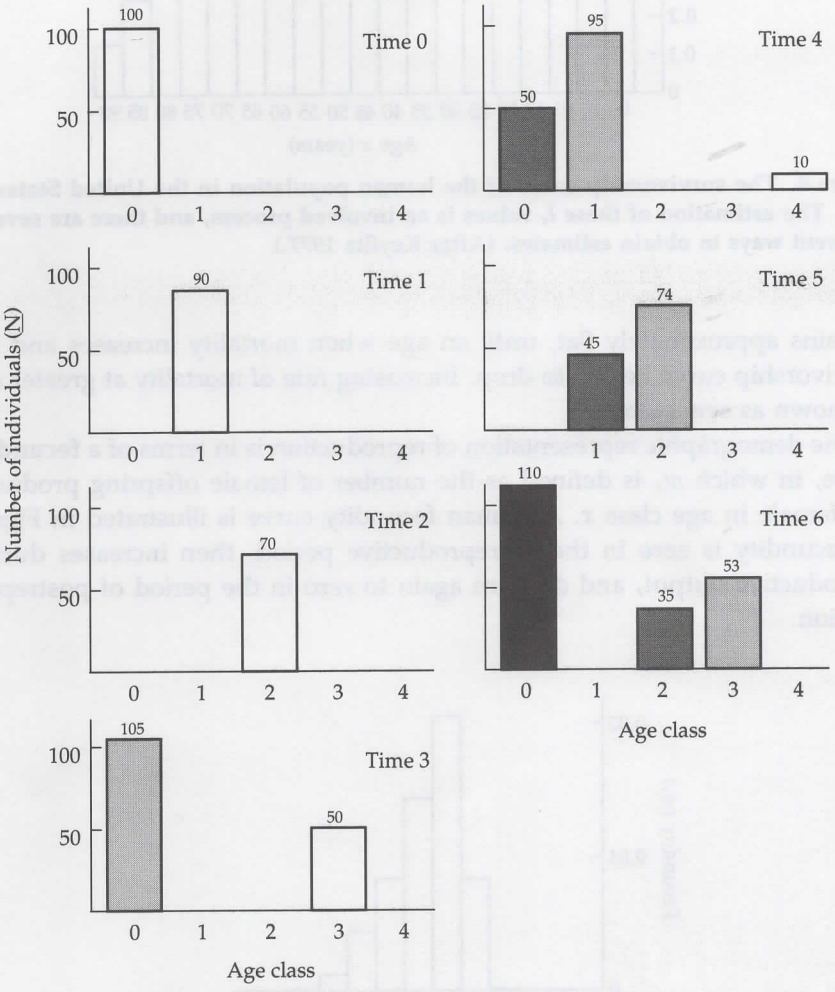


Figure 10. A simple example of a population with four age classes. In this example, the survivorship probabilities are $l_1 = 0.9$, $l_2 = 0.7$, $l_3 = 0.5$, and $l_4 = 0.1$. The fecundities are $m_1 = 0$, $m_2 = 1.5$, $m_3 = 1$ and $m_4 = 0$. There is an initial delay before there are any births into the first age class because there are no individuals of reproductive age until time 2.

a second wave of newborns passes through the population. In time, for most biologically reasonable l_x and m_x distributions, the population settles to a **stable age distribution**, in which each age class makes up a constant proportion of the population. When a population is in stable age distribution, the rate of population growth is given by the l_x and m_x curves according to the Euler equation, which for discrete age classes is

$$\sum_{x=0}^{\infty} e^{-rx} l_x m_x = 1 \qquad \textbf{(9.18)}$$

The value of r that satisfies Equation 9.18 is the **intrinsic rate of population increase**, and the population grows according to

$$N_t = e^{rt} N_o \qquad \textbf{(9.19)}$$

The l_x and m_x curves can also be given as continuous (rather than discrete) functions of time, and in this case, the Euler equation is expressed as

$$\int_{x=0}^{\infty} e^{-rx} l_x m_x dx = 1 \qquad \textbf{(9.20)}$$

Again, the root r of this equation determines the rate of population increase. Note that the population increases in size by a factor of e^r in each time unit, *not* in each generation. The definition of *generation* is much more complex in models with age structure, since successive generations overlap. The effective generation time depends in a complex way on both the l_x and m_x curves.

Selection in Age-Structured Populations

When natural selection operates in populations that are partitioned into age classes, it is necessary to specify the l_x and m_x curves for each genotype. Since fitness is related to the l_x and m_x curves, the outcome of selection depends on these curves. Haldane (1927) and Norton (1928) made initial progress on the problem of selection in age-structured populations, but the problem lay dormant until the 1970s (Anderson and King 1970; Charlesworth 1970, 1972, 1973, 1974, 1980; Pollak and Kempthorne 1970, 1971). A reason for the long dormancy was suggested by Charlesworth (1980, p. 118), who pointed out that many investigators accepted the power and utility of the Malthusian parameter as a measure of fitness. You will see that the formal theory of age-specific selection verifies the utility of the Malthusian parameter, as Fisher (1930) suggested, but it is only accurate for populations in equilibrium with a stable age distribution and unchanging gene frequencies.

Derivation of the general equations for selection in age-structured populations is laborious, but the essentials are straightforward if we assume that the population is in a stable age distribution. A more general derivation can

be found in Charlesworth (1980, pp. 118–152). Consider one gene with two alleles in a diploid population. Genotypes *AA*, *Aa*, and *aa* have survivorship curves $l_{x,11}$, $l_{x,12}$, and $l_{x,22}$ respectively, and they have fecundity curves $m_{x,11}$, $m_{x,12}$, and $m_{x,22}$. In a population consisting solely of genotype *AA*, the population would eventually attain a stable age distribution with an intrinsic rate of increase r_{11} determined from the Euler equation:

$$1 = \int_{x=0}^{\infty} e^{-r_{11}x} l_{x,11} m_{x,11} dx \tag{9.21}$$

Analogous intrinsic rates of increase could be calculated for populations consisting solely of *Aa* (r_{12}) and of *aa* (r_{22}). The fitness of a genotype can be defined from Equation 9.21 as

$$w_{ij}(z) = \int_{x=0}^{\infty} e^{-zx} l_{x,ij} m_{x,ij} \tag{9.22}$$

where r_{ij} is the real root of $w_{ij}(z) = 1$. Equilibrium conditions on the w_{ij} terms and p are

$$pw_{11}(z) + qw_{12}(z) = 1 \tag{9.23a}$$

and

$$pw_{12}(z) + qw_{22}(z) = 1 \tag{9.23b}$$

The root of Equations 9.23 is found by using the constraint $p + q = 1$ and solving the linear system of equations. The solution, first found by Norton (1928), satisfies

$$[1 - w_{12}(z)]^2 = [1 - w_{11}(z)][1 - w_{22}(z)] \tag{9.24}$$

The solution of this equation is $\hat{r}$ — the intrinsic rate of increase of the population at genetic equilibrium and at stable age distribution. The value of $\hat{r}$ is not trivial to find, but the conditions for the existence and stability of a polymorphic equilibrium are surprisingly simple. The conditions are that

$$r_{11} < r_{12} > r_{22}$$

In words, overdominance in the intrinsic rates of increase of the three genotypes is required for stable polymorphism. However, in applying demographic theory, we have assumed that the l_x and m_x curves remain constant. This means that the population at stable age distribution grows without bound at a rate $\bar{r}$. This implicit assumption of density independence gives the r terms the leading role, but if the l_x and m_x curves are density dependent, then their properties would dominate the outcome of natural selection as the carrying capacity is approached.

The Evolution of Senescence

Although natural selection favors genotypes with high intrinsic rates of increase, senescence is nearly universal in higher organisms. Senescence would appear to be maladaptive; nevertheless its prevalence suggests that it has resulted from common evolutionary principles. Application of the principles of selection in age-structured populations to the evolution of senescence provides a theoretical rationale for its occurrence. Figure 11 shows age-specific fecundity and survivorship curves in a hypothetical organism. If a mutation occurred that increased l_x, then the genotype would be favored by selection and r would increase. Similarly, an increase in fecundity would be favored. However, mutations with age-specific effects are usually more complex in that changes early in life have a much greater effect on r. Mutations may affect both early- and late-life characteristics, and they may be pleiotropic in their influences on survivorship and fecundity. Figure 11a shows the effects of a pleiotropic mutation (dashed line) on age-specific survivorship. A mutation that increases early-age survivorship may result in a higher intrinsic rate of increase even if it actually decreases late-age survivorship. In terms of the Euler equation, survivorship beyond the age of reproduction does not affect r. Figure 11b also shows how the survivorship mutation has pleiotropic effects on fecundity, since the increase in early fecundity results in an increase in r. Since r determines whether the mutation can invade, the **antagonistic pleiotropy** resulting from opposing effects of the mutation at early and late ages can sometimes favor senescence mutations that favor early reproduction at the cost of late survivorship.

The idea that senescence can evolve from the decreasing intensity of selection with advancing age was suggested by Medawar (1952). He proposed using Fisher's reproductive value as the measure of relative fitness in age-structured populations. *Reproductive value* is the expected future contribution of offspring over the remaining lifetime of an individual with a prescribed age. Hamilton (1966) formalized the mathematical treatment of the problem and found fault with Medawar's use of reproductive value, suggesting instead measures that are more directly related to changes in gene frequency.

Genetic variation in life history traits (l_x and m_x distributions) has also been characterized in field populations, where ecological constraints can be important. A striking case of life history shift following an ecological perturbation was documented by Hairston and Walton (1986). Normally the copepod *Diaptomus sanguineus* begins diapause in late March in order to avoid predation by fish in the summer. In one area after a severe drought one pond totally dried up, killing all the fish, while larger ponds did not disappear and the fish remained. Over subsequent years, the copepods in the fish-

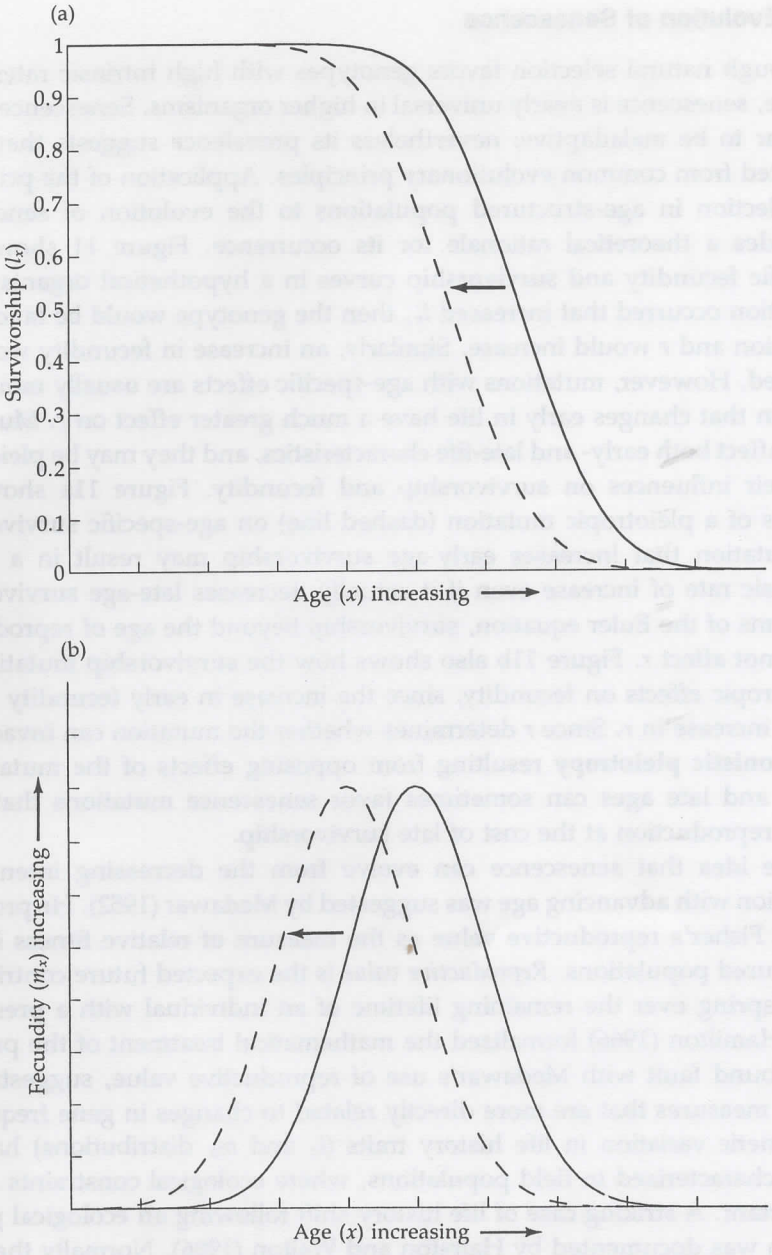


Figure 11. Antagonistic pleiotropy. (a) The mutation (dashed line) results in an earlier mean age of mortality. (b) The same mutation causes an earlier age of reproduction. Under appropriate conditions on the m_x and l_x curves, such joint changes can increase r, in which case they can be favored by natural selection even though they cause earlier death.

depleted pond shifted their time of diapause to nearly one month later in the spring. Although relaxed selection from the absence of fish predation may have resulted in this shift, this remains just a hypothesis in the absence of genetic data. The study does underscore the point that ecological factors such as predation can have a major influence on life history evolution. The importance of predation on life history evolution has been well documented in guppies (Reznick 1982; Reznick and Bryga 1987).

Experimental Studies of Senescence in *Drosophila* C

Because mutations that increase survivorship or fecundity would be expected to increase in frequency to fixation, polymorphism for alleles affecting these traits can only be maintained by antagonistic effects in other fitness components. The existence of antagonistic pleiotropy can be tested by estimating the magnitude of genetic correlation between early- and late-age components of fitness. Antagonistic pleiotropy predicts a negative correlation, meaning that polymorphic genes resulting in senescence would be associated with higher early fecundity (or survivorship). This hypothesis has been tested in *Drosophila* using correlated selection response, sib analysis, and by examination of chromosome substitution lines.

A typical scheme for selection experiments has been to select early reproduction lines by collecting eggs laid by young females, and late reproduction lines by collecting eggs of older females (Rose and Charlesworth 1981b; Luckinbill et al. 1984; Rose 1984; Luckinbill and Clare 1985). A response consistent with antagonistic pleiotropy is shown in Figure 12, in which the line selected for later fecundity also had greater longevity. Similarly, Mueller (1986) maintained lines undergoing natural selection at high and low densities, and observed that the *r*-selected lines had lower fecundity late in life than the *K*-selected lines. He interpreted this difference as consistent with the mutation-accumulation hypothesis, since the *r*-selected lines had relaxed natural selection on late fecundity; presumably mutations with deleterious effects on late fecundity could accumulate since their effect on fitness was negligible.

A second approach to estimating age-specific genetic correlations among fitness components uses full-sib or half-sib designs. Rose and Charlesworth (1981a) studied an artificial population of *Drosophila* and found a negative genetic correlation between early fecundity and survival. Other sib analyses in *Drosophila* found positive genetic correlations between early- and late-age components (Giesel 1979, 1986; Giesel et al. 1982), but these sib designs have been the subject of much debate. Inbreeding in novel environments (the laboratory) could induce an apparent positive genetic correlation, since inbreeding would render deleterious alleles homozygous, whereas in an equilibrium population, the expression of deleterious alleles would result in their elimination by natural selection (Service and Rose 1985). Rose (1984) showed experimentally that inbreeding could generate positive correlations. These laboratory sib studies illustrate a problem in studying the quantitative genetics of natural populations: although samples

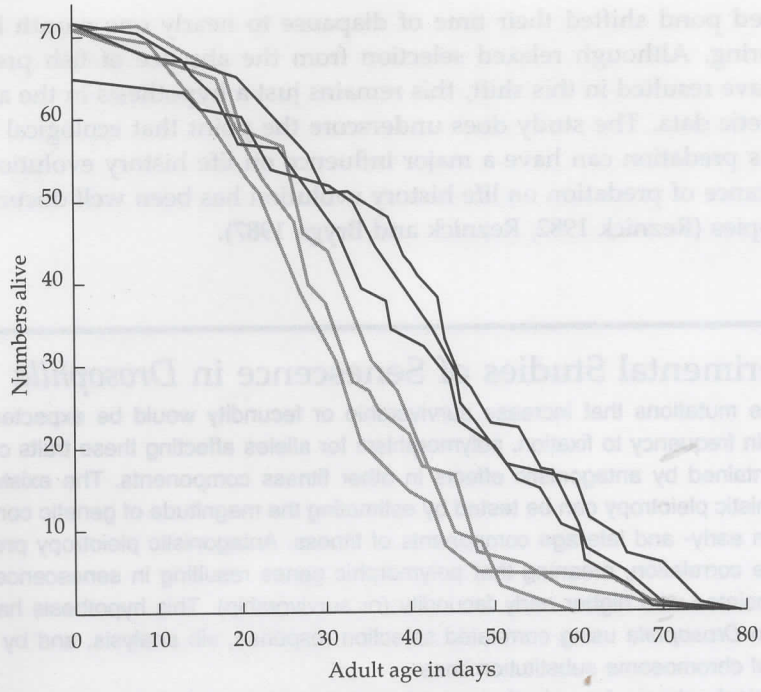


Figure 12. Selecting for later fecundity results in greater longevity in *Drosophila*. The gray lines are survivorship curves of females from control populations, and the black lines are survivorship curves of females from populations that had been selected for later fecundity for 15 generations. (After Rose 1984.)

from natural populations might be representative of the variation found in equilibrium populations in nature, the laboratory environment is novel, so estimates of components of quantitative genetic variance may not be representative (Chapter 8).

The third approach to estimating genetic correlations in life history components uses chromosome substitution or homozygous lines. This method has the appeal that the lines can be used for many other purposes, but unless heterozygotes are tested directly by crossing among lines (diallel design), the components of additive and dominance variance cannot be separated. Hiraizumi (1961, 1985), Simmons et al. (1980), and Hughes and Clark (1988) used extracted lines in various ways to examine genetic correlations and pleiotropic effects of mutations.

PHYSIOLOGICAL GENETICS, ALLOZYMES, AND JUST-SO STORIES

The widespread polymorphism of allozymes led many investigators to attempt to find adaptive reasons for the variation. Allozymes span a diverse range of enzyme functions, including the glycolytic and Krebs cycle enzymes,

those in nucleic acid and amino acid synthesis, and detoxification. In some cases it was possible to postulate a plausible relation between an environmental variable and an enzyme: for example, mice might need a different hemoglobin at high elevations, and mussels that face different osmotic stresses might need different allozymes to mediate osmoregulatory functions. Gould and Lewontin (1979) raised philosophical objections to this adaptationist approach, because there is danger in losing objectivity in trying to identify the adaptive basis of a polymorphism, and adaptationist explanations are limited only by the imagination of the investigator. There is danger in *post hoc* reasoning—in assuming that there *must* be an adaptive basis to a polymorphism, then performing studies and statistical tests until something comes out significant, then finally rationalizing the significant result according to some unverified chain of causality (called a "just so" story, after Kipling's charming book of adaptationist tales for youngsters). Although loss of objectivity was a problem with some studies that sought physiological reasons for allozyme polymorphisms, it is not impossible to proceed objectively. For example, modern molecular methods that allow inference of identity of alleles provide very strong tests of neutrality. Rejection of neutrality justifies a search for the forces that impinge upon the polymorphism.

Allozymes and Metabolic Control

Since enzymes catalyze reactions that convert substrates into products, and allozymes are altered forms of enzymes, any effect of allozymes on fitness must be mediated by altered kinetics of the reactions they catalyze. Although theories of enzyme kinetics for single reactions are well developed, starting with Michaelis and Menten (1913) and Briggs and Haldane (1925), only recently has the theory been extended to networks of many enzymes in complex metabolic pathways. This theory, known as **metabolic control theory**, provides a needed translation from enzyme kinetics to fitness, since fitness is not determined by single reactions but by the aggregate effects of interconnected metabolic pathways. The following gives a summary of the essentials of metabolic control theory as it has so far been applied in population genetics.

A convenient starting point is a linear metabolic pathway with successive substrates and products S_i related as follows:

$$S_1 \rightleftharpoons S_2 \rightleftharpoons S_3 \rightleftharpoons \ldots S_n \rightleftharpoons S_{n+1} \tag{9.25}$$

At the risk of some confusion we will use the symbols S_i to refer to the concentrations of the substrates as well as to the substrates themselves. A classical equation due to Briggs and Haldane (1925) gives the velocity v_i of a

reaction between a substrate with concentration S_i and a product with concentration S_{i+1} as follows:

$$v_i = \frac{(V_i/K_i)(S_i - S_{i+1}/Q_{i,i+1})}{1 + S_i/K_i + S_{i+1}/K_i'} \tag{9.26}$$

In Equation 9.26, V_i is the maximum velocity (often called V_{max}) and K_i the Michaelis constant (often called K_m) of the forward reaction ($S_i \rightarrow S_{i+1}$); K_i' is the Michaelis constant of the reverse reaction ($S_{i+1} \rightarrow S_i$); and $Q_{i,i+1}$ is the equilibrium constant across the step. In kinetic terms, V_i is the velocity that the reaction approaches when $S_i \rightarrow \infty$, K_i is the value of S_i at which the reaction rate is $(1/2)V_i$ (assuming $S_{i+1} \approx 0$), and $Q_{i,i+1}$ is the ratio of S_{i+1}/S_i at equilibrium when $v_i = 0$.

When the enzymes are unsaturated ($S_i << K_i$, $S_{i+1} << K_i'$), Equation 9.26 can be approximated by

$$v_i = (V_i/K_i)(S_i - S_{i+1}/Q_{i,i+1}) \tag{9.27}$$

Equation 9.27 specifies a set of simultaneous equations for the velocities of the reactions in the pathway. While there is no general solution for the equations, Kacser and Burns (1973) showed how they could be analyzed in the case when the pathway is in a steady state. In a steady state, all reaction rates must be equal to each other, and therefore

$$v_i = J \tag{9.28}$$

where J is the flux through the pathway, and $i = 2, 3, \ldots, n - 1$. With these assumptions, Kacser and Burns (1973) showed the relation between the flux and the parameters of the enzymes is

$$J = \frac{S_1 - S_{n+1}/Q_{1,n+1}}{\sum_{i=1}^{n} \frac{K_i}{V_i Q_{1,i}}} \tag{9.29}$$

where $Q_{1,i}$ is the product of the equilibrium constants of all the reactions between S_1 and S_i ($Q_{1,i} = Q_{1,2} \times Q_{2,3} \times \ldots \times Q_{i-1,i}$). If the first and final substrate concentrations (S_1 and S_{n+1}) are held constant, we can express flux as a function of the activities E_i of the enzymes by defining $E_i = V_i Q_{1,i}/K_i$, and then

$$J = \frac{C}{(1/E_1) + (1/E_2) + \ldots + (1/E_n)} \tag{9.30}$$

where C is a constant ($C = S_1 - S_{n+1}/Q_{1,n+1}$) and E_i is the activity of enzyme i.

Equation 9.30 assumes that the initial substrate S_1 and the final product S_{n+1} are held at fixed concentrations in the cell by factors extrinsic to the

pathway, so changes in flux J do not affect S_1 and S_{n+1}. Furthermore, Equation 9.30 assumes Briggs-Haldane kinetics with unsaturated enzymes (Equation 9.27), which implies that substrate and enzyme molecules exist in a homogeneous solution and their probability of colliding is the product of their concentrations. If there is cellular compartmentalization or the enzymes exist in a multienzyme complex, then the equations for flux become more complex. However, Kacser and Porteous (1987) have argued that the qualitative conclusions of metabolic control theory are actually more general than the restrictive assumptions of the present derivation might imply.

Kacser and Burns (1981) used an equation like Equation 9.30 to suggest why most mutations might be expected to be recessive to the wildtype counterparts. Their argument provides a possible molecular explanation for the widespread phenomenon of dominance. If most mutations occur in enzymes that are embedded in pathways like the one in Equation 9.25, and the phenotypic effects result from changes in pathway flux, then Equation 9.30 implies that the marginal effects of small changes in enzyme activity can be very small. That is, if the E_i terms in Equation 9.30 are all much greater than zero, then small increases or decreases in activity will have even smaller effects on flux. Moreover, the longer the pathway, and the more evenly distributed the activities of the enzymes, the smaller the effect of a small change in the activity of any one enzyme on the flux. The expected relation between flux and the activity of an enzyme embedded in a linear pathway is illustrated in Figure 13. Kacser and Burns (1981) document cases in which the wildtype enzyme is so active that the flux is well out on the plateau portion of the hyperbolic curve in Figure 13. In such cases, heterozygotes for null alleles, which have approximately half the wildtype enzyme activity, still have nearly a wildtype flux through the pathway. This disproportionately small effect on flux is one molecular basis of dominance of wildtype alleles.

In some cases, the fitness of an organism is proportional to flux across a metabolic pathway, so the relation between fitness and enzyme activity is also hyperbolic. For example, in *Escherichia coli* chemostats in which growth is limited by shortage of lactose, small changes in β-galactosidase activity result in even smaller changes in growth rate (Dean et al. 1986). Similarly, the relation between fitness and permease activity is hyperbolic in lactose-limited chemostats (Dykhuizen et al. 1987). In situations in which flux through metabolic pathways is closely related to fitness, selection is directional and favors increased flux. This inevitably results in increasing enzyme activity until a point of diminishing return is reached in which further activity increases result in negligible increases in flux. At this point the effect on fitness of mutations that change enzyme activity by a small amount will be so small that $Ns << 1$ and random drift will be the dominant force affecting

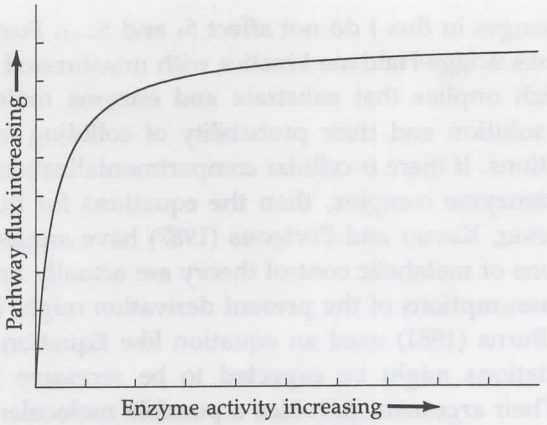


Figure 13. The relationship between the metabolic flux through a biochemical pathway and the activity of one of the enzymes in that pathway. The relationship is that of Equation 9.30 for a hypothetical set of parameters. Note that beyond a certain activity of the enzyme, the flux is relatively constant, implying that other aspects of the pathway are limiting the flux.

the allele frequencies of these mutations. In this sense, neutrality is the outcome of selection (Hartl et al. 1985).

The Kacser and Burns model should not be construed as a universal explanation for all allozyme polymorphisms. Burton and Place (1986) emphasize a number of limitations of the theory, including the equating of flux with fitness, the possible affects of metabolic pool sizes on fitness, the greater sensitivity of some enzymes at metabolic branch points, and the importance of homeostasis in a changing environment (for rebuttal see Hartl et al. 1986). Extensions of metabolic control theory that account for nonsteady-state conditions (Easterby 1973; Heinrich and Rapoport 1983) and branch point effects (LaPorte et al. 1984) are actively being developed.

A generalization of the explanation of dominance in terms of metabolic control theory is that pleiotropic effects as well as dominance might be mediated by effects of enzyme variants on flux (Keightley and Kacser 1987). This approach has the potential to unite metabolic control theory with quantitative genetics, since quantitative variation for enzyme activities is readily detected (Laurie-Ahlberg et al. 1980, 1982). Experimental systems that relate a physiological phenotype to enzyme kinetics, such as flight metabolism to α-glycerophosphate dehydrogenase activity (Barnes and Laurie-Ahlberg 1986), lipid storage to fatty acid synthetase activity (Geer et al. 1986; Clark and Keith 1988), and the chemostat studies described above, are the experimental counterparts of metabolic control theory.

Population Genetics of Phenylketonuria

D

The clinical consequences of altered forms of key enzymes are painfully evident in humans, and they constitute a class of genetic disorders known as **inborn errors of metabolism**. However, understanding the molecular and clinical basis of a disorder does not explain its population genetics. Alleles that result in dysfunctional phenylalanine hydroxylase and cause phenylketonuria (PKU) serve as examples. PKU is inherited as a simple autosomal recessive. Homozygosity for a defective allele results in loss of phenylalanine hydroxylase activity, which leads to severe mental retardation unless detected at birth and counteracted with a strict diet low in phenylalanine. However, molecular analysis has shown that there is more than one mutant form of the allele, and that they lead to loss of function by several routes. Moreover, the molecular methods are sufficiently detailed that particular alleles can be inferred to be identical by descent. For example, about 38% of PKU alleles in a Danish sample have identical high-resolution restriction maps. This allele, known as haplotype-3, features an abnormal intron–exon splice junction (DiLella et al. 1986). Another haplotype, haplotype-2, accounts for 20% of the PKU alleles and is characterized by a single amino acid substitution (DiLella et al. 1986). Given all the possible ways to lose enzyme function, it is noteworthy that these two alleles account for over half of the defective alleles in the population. Kidd (1987) has argued that such an occurrence is unlikely from random drift alone, and that there may be selection, perhaps favoring heterozygotes, that enabled these two haplotypes to attain high frequency. Although no such selective mechanism has been proven, this example shows that even a genetic disorder with an obvious effect in homozygotes may have unobserved effects in heterozygotes that are important to the population dynamics.

Ecological Genetics of Single-Gene Substitutions

Allozyme polymorphisms with pronounced geographic patterns in allele frequency have attracted much attention as potentially favorable cases for identifying ecological mechanisms of selection. The idea that geographical patterns might be determined by fitness differences of genotypes for single genes might seem unlikely, since the marginal effect of a single gene on a complex quantitative character like fitness must usually be small. Even if geographical patterns do result from different kinetic properties of the allozymes (making genotypes differently adapted to the environments), the demonstration of fitness differences based on kinetic properties of allozymes is difficult. However, several well-studied human genetic disorders show that single-gene substitutions can have a major effect on fitness. For example, the relation between malaria and sickle-cell anemia inspires the search for simple selective mechanisms underlying geographical patterns in other organisms.

A striking **cline**, or geographic gradient, in the frequency of alleles of the *aminopeptidase-1* locus (*LAP*) in the marine mussel *Mytilus edulis* is found in Long Island Sound (Figure 14). Koehn and associates have characterized many of the physiological effects that relate this polymorphism to ecological conditions in the sound (reviewed by Koehn and Hilbish 1987; Hilbish and Koehn 1985). *M. edulis*, like most marine invertebrates, regulates cell volume by remaining isosmotic with the surrounding seawater. When the concentration of salt in the sea water varies, as it does in Long Island Sound, the osmolarity within cells must be adjusted accordingly. Intracellular osmolarity is increased by aminopeptidase proteolysis of cellular proteins to free amino acids, and osmolarity is decreased by excretion. Genotypes with the lap^{94} allele have a higher level of aminopeptidase activity than do other genotypes, and this is the allele that is present in highest frequency in the high-salinity oceanic populations. The higher activity is also reflected in a higher rate of accumulation of free amino acids when lap^{94} individuals are placed in high-salinity saltwater. Genotypes bearing the lap^{94} allele also excrete nitrogen at a higher rate, and in the period from September to December, lap^{94} genotypes in the brackish parts of Long Island Sound can lose an excessive amount of nitrogen. Higher mortality of lap^{94} genotypes at this time of the year has been well documented (Koehn et al. 1980). The lap^{94} allele appears to be dominant to the other two common alleles with respect to all physiological parameters measured. Although there are still some uncertainties concerning the mechanism of the pleiotropic effects of the *LAP* polymorphism, the ability to perform direct manipulative physiological tests in the laboratory has been of great value in understanding the basic biology of this system.

The importance of osmoregulation mediated by the enzyme glutamate-pyruvate transaminase was shown in the intertidal marine copepod *Tigriopus californicus* through its influence on levels of intracellular alanine (Burton and Feldman 1983). Genotypes with higher activity forms of this enzyme produce intracellular alanine at a higher rate, and survive better under osmotic stress.

Another striking allozyme cline is found in allele frequencies at the *LDH-B* locus in the marine fish *Fundulus heteroclitus* along the east coast of the United States (Powers et al. 1983). The important physiological variable along this cline is temperature, where a one-degree increase in latitude corresponds to a 1°C decrease in mean temperature. Lactate dehydrogenase catalyzes the reaction from lactate to pyruvate, which is important in avoiding lactic acidosis during heavy exercise. At 10°C the genotype *LDH-*B^bB^b has a higher activity than *LDH-*B^aB^a for the reaction in both directions (lactate oxidation and pyruvate reduction), but at 25°C there is no difference. The heterozygotes have intermediate activities. When fish of the homozygous genotypes are tested in a swimming chamber at 10°C (Figure 15), the B^bB^b fish can sustain a 20% higher swimming speed than B^aB^a fish. This stamina

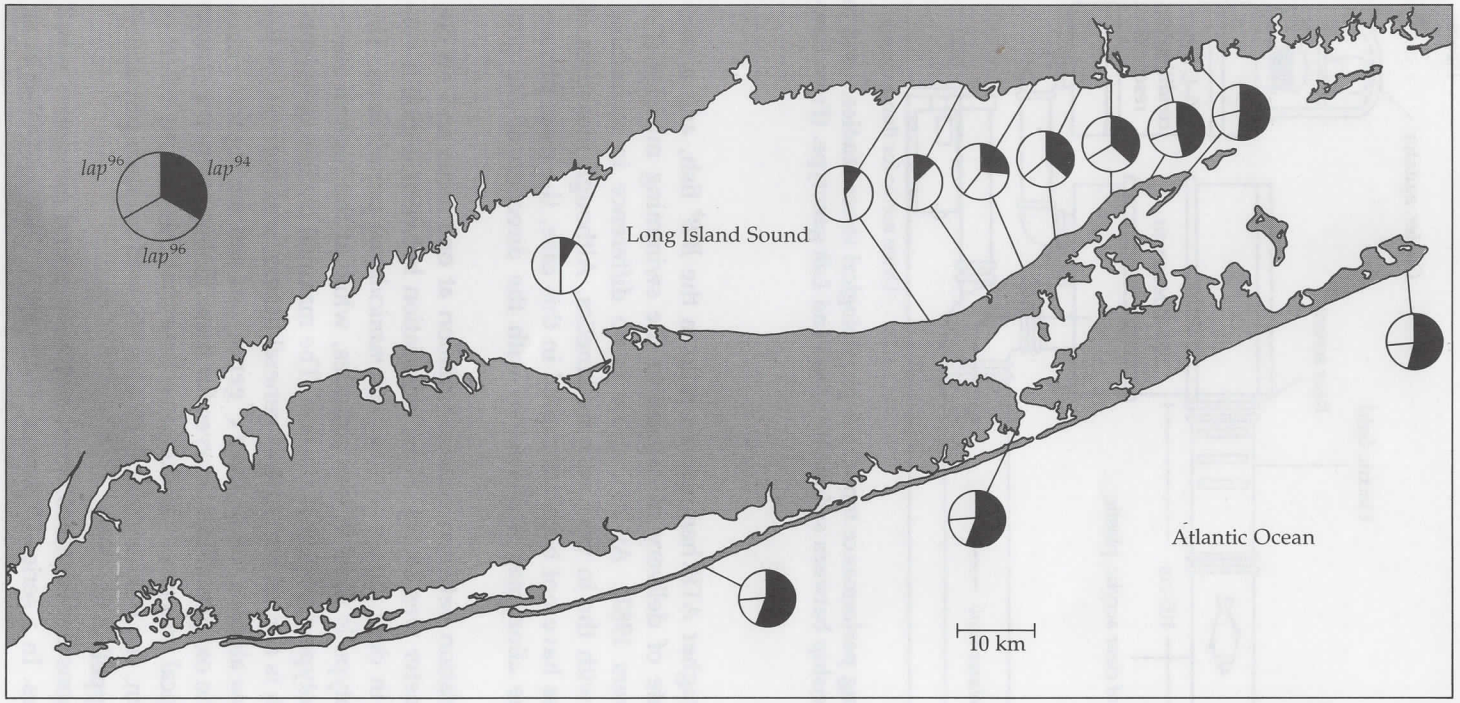


Figure 14. The frequency cline of the *lap*^{94} allele in the mussel *Mytilus edulis*. This allele is much more frequent on the Atlantic coast than it is within Long Island Sound. (After Koehn and Hilbish 1987.)

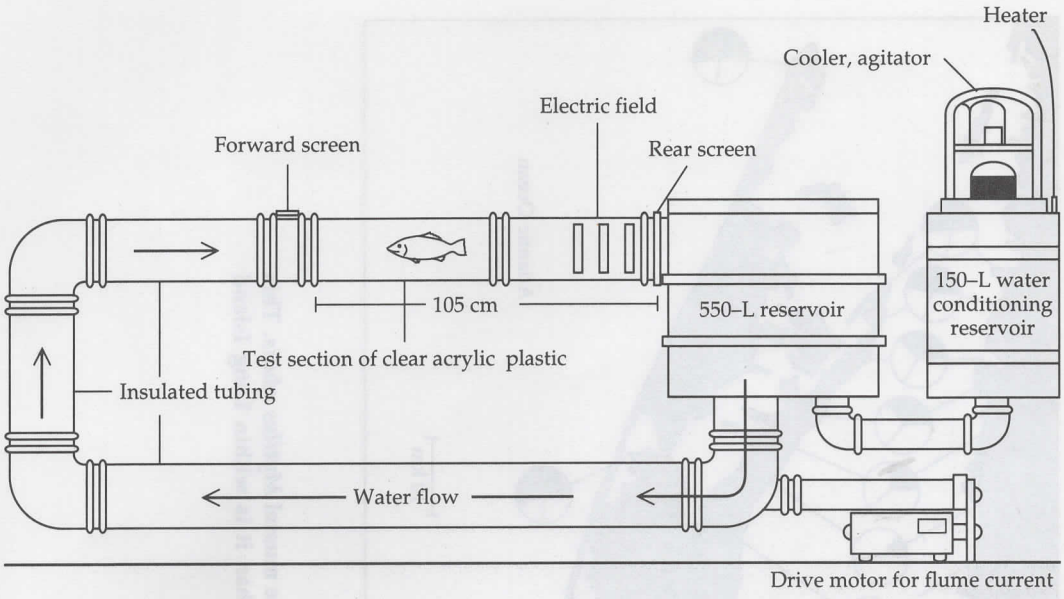


Figure 15. Swimming performance tunnel for physiological stress studies, used to establish the relationship between swimming speed and *Ldh* genotype. (From Powers et al. 1983.)

is mediated by a higher ATP:hemoglobin ratio in the B^bB^b fish, and a consequent greater rate of delivery of oxygen to the swimming muscles (DiMichelle and Powers 1982). At 25°C there is no difference in swimming ability, consistent with the in vitro enzyme kinetics. Although measures of reproductive fitness have not been attempted in this case, the clear physiological effect of the allozymes is consistent with the direction of the geographic cline.

A striking correlation between genetic variation at one locus and an environmental parameter was found in the association between α-globin variants and altitude in deer mouse (*Peromyscus maniculatus*) populations. The a^1c^1 α-globin haplotype is fixed at low altitudes, while at the highest elevations, the a^0c^0 haplotype approaches fixation. The maximal oxygen consumption (V_{O_2max}), which is a good measure of general fitness, was highest for the a^1c^1 genotypes at low altitude, while the a^0c^0 genotypes had the highest V_{O_2max} at high altitude. The oxygen affinity curves for these globins were consistent with the physiological measures. At the same time that this striking effect of α-globin was shown, the polymorphic β-globins showed no such physiological variation (Chappell et al. 1988).

Not all associations between ecological adaptations and allozymes result in geographic clines. In a series of papers, Watt (1977, 1983) and Watt et al.

(1983, 1985) examined the biochemical, physiological, and ecological consequences of polymorphisms at the *phosphoglucose isomerase* (*PGI*) locus of the butterfly *Colias philodice eriphyle*. This system is more complex than the preceding in that it is multiallelic, and the population consists of a series of cohorts each flight season. Detailed kinetic tests were performed on the enzyme from genotypes PGI^3/PGI^3, PGI^3/PGI^4 and PGI^4/PGI^4, and the heterozygotes were found to have the highest catalytic efficiency at low temperature. This finding was correlated with the heterozygote advantage in the time of flight initiation in the morning and the overall flight density. Flight is thought to be a critical component of fitness since it is important in feeding, mating, and oviposition. Direct estimates of mating success also support the PGI^3/PGI^4 advantage.

Tolerance to desiccation is another ecological factor that has been implicated in certain polymorphisms. Templeton et al. (1985) observed a polymorphism in *Drosophila mercatorum* associated with a phenotype *abnormal abdomen* (*aa*), which is characterized by a juvenile form of the cuticle that persists on the adult abdomen. Molecular analysis revealed that the 28S coding sequence of ribosomal DNA was interrupted by a 5–6-kb insert in *aa* flies (DeSalle et al. 1986). The insertion is thought to interrupt function of the rDNA at a critical time in development, and when there are insufficient copies of the normal rDNA, fewer functional ribosomes are produced and protein production is limited. In *D. mercatorum*, protein synthesis is critical at the stage in development when the enzyme juvenile hormone esterase is made; *aa* flies have low titers of this esterase, which accounts for the retention of the juvenile cuticle. In this example the apparently deleterious mutation is maintained in the population by its increased tolerance to desiccation.

In addition to single genes that are involved in local differentiation, combinations of genes may be important in adaptations to local environments. An example is found in the marine copepod *Tigriopus californicus*, which lives in rocky intertidal pools along the Pacific coast of North America and shows marked local differentiation. Protein electrophoresis shows that, within a rocky outcrop, allele frequencies do not differ markedly among intertidal pools (suggesting population mixing); while neighboring rocky outcrops, though separated by less than a kilometer of sandy beach, can be fixed for alternate alleles (Burton and Feldman 1981). Apparently the combination of founder effects and bottlenecks in an unstable environment has resulted in a high level of differentiation among rocky outcrops, and nearly all of 35 outcrops studied show electrophoretically distinct profiles. As for local adaptation, the most direct means of assessing viability differences is to perform crosses and to compare the observed ratios of genotypes among the progeny to the expected Mendelian ratios. Crosses among several pairs of populations were analyzed in this way, and malic enzyme (*Me*) and glutamate oxaloace-

tate transaminase (*Got*) were assayed in the F_2 generation (Burton 1987). The counts of the two-locus genotypes deviated strongly from the Mendelian expectations, and estimates of viability were as low as 1% of the double heterozygote viability. Although these estimates apply to entire chromosomal regions marked by the allozymes, the magnitude of the viability effects nevertheless suggest an impressive degree of differentiation among the populations. In addition, the two-locus viability estimates suggested a degree of nonmultiplicative interaction between the two chromosomal regions.

Most of the above examples make a mechanistic connection between viability and an allozyme, but allozyme variation may also influence fertility. One example concerns esterase-6 and mating behavior in *Drosophila melanogaster*. Esterase-6 is concentrated in the anterior ejaculatory duct in males and transferred as a component of seminal fluid to females during mating. The influence of this enzyme on remating time of females has been demonstrated by the observation that females remate sooner after mating with males having a higher esterase-6 activity, and lines selected for reduced remating time had higher esterase-6 activity, but the ecological implications of this control of remating time remain unclear (Richmond et al. 1980).

Heavy Metal Tolerance in Plants

At various places throughout the world, soil is contaminated by toxic concentrations of one or more heavy metals. Such areas can result from the rare presence of undisturbed metal ore near the soil surface; more commonly soil is contaminated by unproductive ore deposits, tailings, or seepage near actual mining areas. In addition to heavy metal contamination, such soils are typically low in the essential nutrients nitrogen, phosphorous, and potassium. Because of their relative lack of vegetation, these areas generally have soil of poor texture and are lacking in organic matter. Any plants that do survive are subject to exposed, often sand-dunelike conditions (Figure 16).

Inhospitable as such contaminated sites are, some organisms have evolved tolerance not only to the generally poor soil conditions, but also to the toxic metals (Jain and Bradshaw 1966; Antonovics et al. 1971). Certain bacteria and fungi are frequently found in such areas, as are some algae, mosses, lichens, liverworts, and flowering plants; however, ferns and gymnosperms tolerant of heavy metals are rarely found. Mechanisms of resistance to heavy metal poisoning vary among species, ranging from those in which the metal is not taken up by the plant to those in which the metal is taken up but is sequestered in the cell walls, away from vital metabolic processes. Tolerance to heavy metals seems to evolve very rapidly, as mining sites only 50 to 100 years old often have well-established populations (Antonovics et al. 1971).

Figure 16. Desolate appearance of a site contaminated with a heavy metal due to mining operations. Resistant grasses are emerging in the foreground. (Courtesy of J. Antonovics.)

Ecological Genetics of Mycophagous *Drosophila*

E

Animals differ from plants in that behavior is an important component controlling their spatial distribution. To the extent that animals possess genetic variation that allows them to utilize different substrates in the environment, the effect of this variation on fitness is greatly compounded when combined with behaviors for seeking out environments in which the genotypes are best adapted. For example, *Drosophila tripunctata* is a polyphagous fly that feeds on a variety of decaying plants throughout the eastern United States. Jaenike (1985b) performed experiments to elucidate genetic variation in host preference. Flies of two different strains were reared on mushrooms, then fed for one week on either mushrooms, tomatoes, or regular *Drosophila* medium before being field-tested. The flies were released in the center of concentric circles of mushroom and tomato baits, and the data consisted of counts of flies recovered on the two baits. Females fed on tomatoes had a much lower probability of being recaptured on mushrooms, indicating an effect of the previous experience on choice of host. In addition to the effect of experience, there was a remarkable difference between the strains in that females of one strain were recaptured on mushrooms nearly 70% of the time, while females of the other strain were recaptured on mushrooms about 15% of the time. The F_2 generation from crosses of the strains was intermediate in degree of preference for mushrooms. Tests of oviposition showed that, whereas the parental strains had an oviposition preference which was the opposite of their feeding preference, there was no such association in the F_2.

Drosophila tripunctata, *D. putrida* and *D. recens* are all able to undergo development on poisonous *Amanita* mushrooms. These mushrooms contain levels of α-amanitin (a bicyclic octapeptide that inhibits RNA polymerase II) that is lethal to most eukaryotes, including nonmycophagous *Drosophila*. How mycophagous *Drosophila* resist the toxic effects of α-amanitin is not clear, but their RNA polymerase II enzyme is just as sensitive as in other species of *Drosophila*. The lesson here is that organisms are biochemically complex, and even the most seemingly obvious mechanism of adaptation may be incorrect. As for why these *Drosophila* should have evolved resistance to the toxin, it appears that the level of parasitism by nematodes is reduced on *Amanita* mushrooms (Jaenike 1985a), resulting in improved male mating success (Jaenike 1988).

Populations that have evolved heavy metal tolerance often exist side by side with nontolerant populations of the same species, and gene exchange occurs. Although gene exchange is often considered a cohesive force that acts to prevent genetic divergence of populations, the example of heavy metal tolerance shows that selection can nevertheless overcome the effects of migration. The situation has been particularly well studied in certain grasses, among them *Agrostis tenuis* (bent grass), *Agrostis stolonifera* (creeping bent), and *Anthoxanthum odoratum* (sweet vernal grass). Figure 17a shows how an index of lead tolerance in *A. tenuis* changes abruptly across a mine boundary from values above 60 to ones below 40; Figure 17b shows a comparable abrupt change in copper tolerance in *A. tenuis* at another mine; and Figure 17c shows how zinc tolerance in *A. odoratum* at yet another mine declines dramatically from a value of nearly 80 to one of 15 over a distance of only about four meters across the mine boundary. Thus, in spite of gene flow, genetic differentiation associated with heavy metal tolerance at mining sites is sharp and distinct. The genetic basis of heavy metal tolerance is not known with precision, but the trait seems to be polygenic and to have a high broad-sense heritability (Antonovics et al. 1971; MacNair 1979). Moreover, tolerance to one heavy metal does not confer resistance to other heavy metals.

Selection coefficients pertaining to heavy metal tolerance in natural habitats are not known, but they have been studied under experimental conditions simulating natural habitats (Jain and Bradshaw 1966; Cook et al. 1972). The intensity of selection involved can be very large, as shown by the data summarized in Table 1. The fourth column in Table 1 gives estimated fitnesses of nontolerant plants relative to tolerant ones when grown in contaminated soil. With respect to heavy metal tolerance (examples 1 to 3 in the table), selection against nontolerant plants in contaminated soil is very intense, with relative survival rates of nonadapted plants ranging from 1% to 5%. The last column of Table 1 gives estimated fitnesses in uncontaminated soil. Selection in this case is generally less intense than that observed in contaminated soil, with relative fitnesses in the case of heavy metal tolerance ranging from 0.60

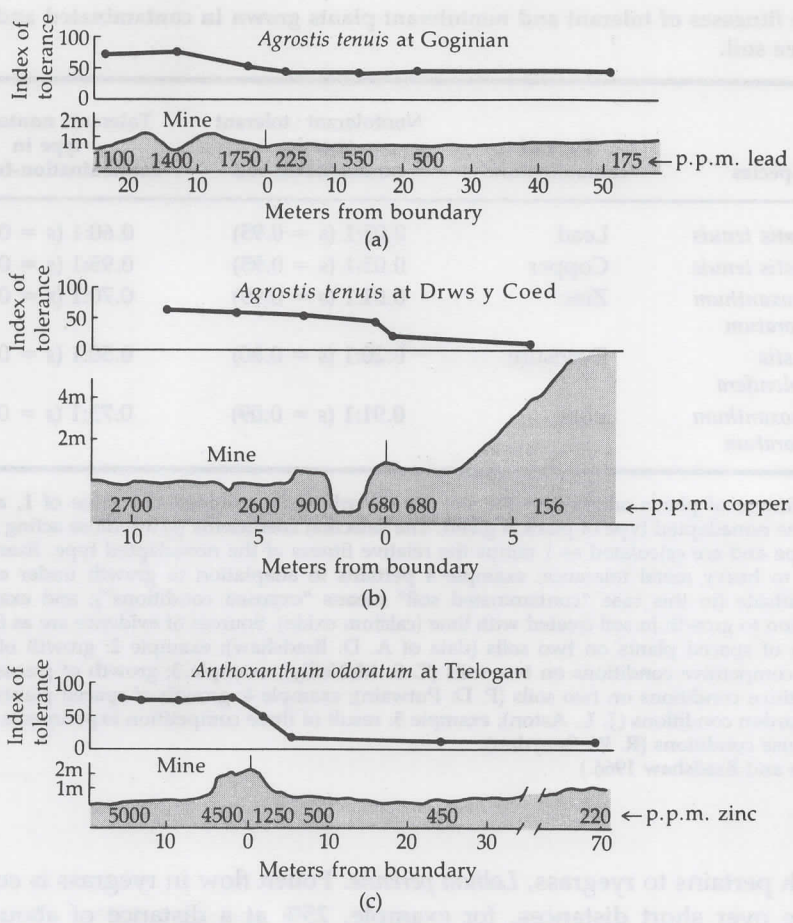


Figure 17. Profiles of three mine sites in Wales showing (below each profile) the amount of heavy metal contamination of the soil and (above each profile) the observed cline of heavy metal tolerance in grasses collected along the profile. Note the particularly abrupt drops in resistance at the mine boundaries in (b) and (c). (After Jain and Bradshaw 1966.)

to 0.95, and the selection favors nontolerant plants. The last two examples of Table 1 give relative fitnesses of plants adapted to weather exposure on a cliffside (example 4) or to limed soil (example 5) rather than to heavy metals. Again the selection coefficients are quite large.

The amount of migration occurring between adjacent tolerant and nontolerant populations is not known (for a review of gene flow in seed plants, see Levin and Kerster 1974). The grasses are wind pollinated, so pollen flow is undoubtedly more important for gene exchange than is seed transport. The pattern of pollen flow is probably similar to that depicted in Figure 18,

Table 1. Relative fitnesses of tolerant and nontolerant plants grown in contaminated and contamination-free soil.

Example	Species	Type of "contamination"	Nontolerant : tolerant type in contaminated soil	Tolerant : nontolerant type in contamination-free soil
1	*Agrostis tenuis*	Lead	0.05:1 (s = 0.95)	0.60:1 (s = 0.40)
2	*Agrostis tenuis*	Copper	0.05:1 (s = 0.95)	0.95:1 (s = 0.05)
3	*Anthoxanthum odoratum*	Zinc	0.01:1 (s = 0.99)	0.70:1 (s = 0.30)
4	*Agrostis stolonifera*	Exposure	0.20:1 (s = 0.80)	0.50:1 (s = 0.50)
5	*Anthoxanthum odoratum*	Lime	0.91:1 (s = 0.09)	0.73:1 (s = 0.27)

In each case, the fitness of plants adapted to the soil type is arbitrarily assigned the value of 1, and the relative fitness of the nonadapted type of plant is given. The selection coefficients (s) are those acting against the nonadapted type and are calculated as 1 minus the relative fitness of the nonadapted type. Examples 1 through 3 pertain to heavy metal tolerance; example 4 pertains to adaptation to growth under exposed conditions on a cliffside (in this case "contaminated soil" means "exposed conditions"); and example 5 pertains to adaptation to growth in soil treated with lime (calcium oxide). Sources of evidence are as follows. Example 1: growth of spaced plants on two soils (data of A. D. Bradshaw); example 2: growth of plants under spaced and competitive conditions on two soils (T. S. McNeilly); example 3: growth of plants under spaced and competitive conditions on two soils (P. D. Putwain); example 4: growth of spaced plants under exposed cliff and garden conditions (J. L. Aston); example 5: result of three competition experiments carried out under greenhouse conditions (R. W. Snaydon).
(Adapted from Jain and Bradshaw 1966.)

which pertains to ryegrass, *Lolium perenne*. Pollen flow in ryegrass is considerable over short distances, for example, 25% at a distance of about one meter. The band in Figure 18 falls off rapidly, and the amount of pollen flow at distances of five meters or more is down to roughly 10%. On the other hand, genetic differentiation between heavy metal tolerant and nontolerant populations at mine boundaries occurs over distances of only a few meters, so it is clear that the genetic differentiation is maintained in the face of substantial gene exchange.

As noted earlier, abandoned mine sites differ from surrounding areas not only in their heavy metal contamination but also in other factors such as lower nutrient levels and higher degrees of exposure. Natural selection for heavy metal tolerance is therefore only one aspect of evolutionary adaptation on abandoned mine sites. For *Anthoxanthum odoratum* at the Trelogan mine site, clines for such traits as plant height and flowering time occur (Antonovics and Bradshaw 1970), with plants at the mine site being about 30% shorter (presumably an adaptation to exposure) and flowering several days earlier (which tends to reduce contamination by pollen flow from nontolerant

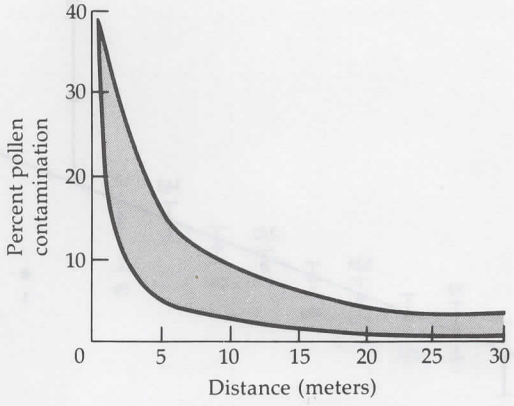


Figure 18. Gene flow from a source as measured by the amounts of genetically marked pollen in perennial ryegrass *Lolium perenne*, a wind-pollinated species. The curve is a gray band rather than a solid line because the amount of gene flow from a source depends on direction relative to the wind. (After Griffiths 1950.)

plants). A remarkable further finding is that *A. odoratum* at the Trelogan mine undergoes self-fertilization about 1% of the time, in contrast to a virtual absence of self-fertilization in nearby tolerant plants; this suggests selection for reproductive isolation (Antonovics 1968; Caisse and Antonovics 1978).

HETEROZYGOSITY AND GROWTH RATE

Several population surveys have revealed a correlation between the number of heterozygous loci that an individual has and growth rate — it appears more highly heterozygous individuals grow faster. This correlation has been observed in oysters (Zouros et al. 1980), mussels (Koehn and Gaffney 1984), coot clam (Koehn et al. 1988), pitch pine (Ledig et al. 1983), tiger salamanders (Mitton et al. 1986), and other organisms, although it is not universal (Mitton and Grant 1984). The study by Koehn et al. (1988) of the coot clam *Mulinia lateralis* is the largest to date, with 1906 individuals assayed for 15 allozyme genes and four measures of growth made. Clam larvae, collected by dredging, were measured and placed in individual compartments in sea water. After two weeks of feeding, the animals were frozen, measured, and typed electrophoretically. Figure 19 shows the relationship between the number of heterozygous genes and the initial shell length. The correlation coefficient is $\langle r \rangle = 0.255$ ($P < 0.001$). Multiple regression showed that enzymes for protein catabolism and glycolysis had the strongest relation to growth rate, while other enzymes, such as those in the pentose shunt, were not correlated.

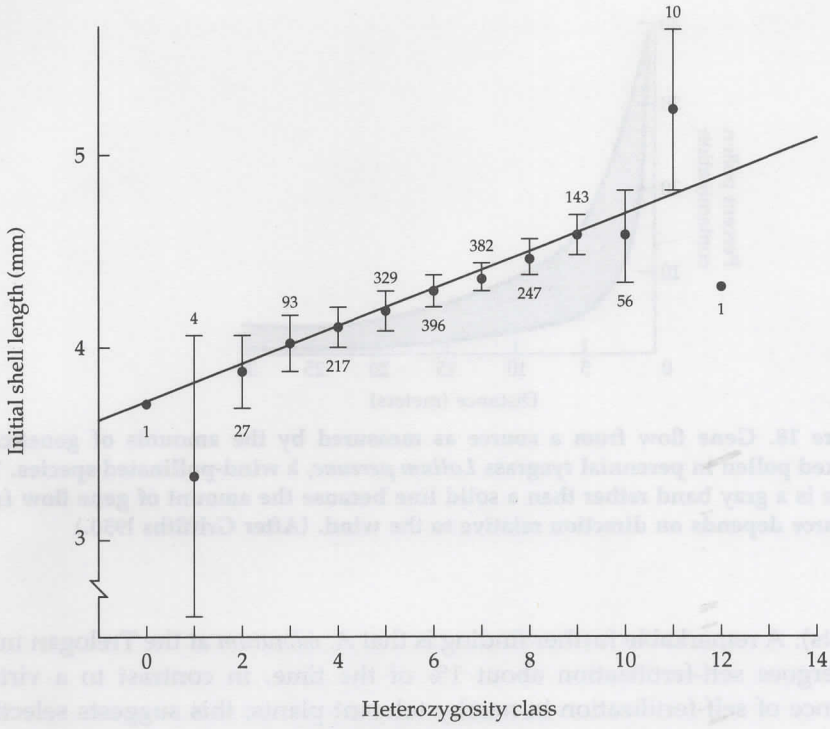


Figure 19. The relationship between individual heterozygosity, based on 15 loci, and initial shell length among 1906 individuals of *Mulinia lateralis*. The average initial shell length and associated 95% confidence interval is given for each heterozygosity class. Numbers at each heterozygosity class are the sample sizes. The two variables are significantly correlated ($\langle r \rangle = 0.255$, $P < 0.001$) among individuals. (From Koehn et al. 1988.)

How can these results be explained? From a theoretical point of view, if multilocus equilibrium is maintained by selection, then in the absence of linkage disequilibrium, the average fitness always increases with the number of heterozygous loci (Turelli and Ginzburg 1983). Numerical simulations suggest that fitness increases with the number of heterozygous loci in most models in which the population is at a multilocus equilibrium maintained by selection. However, the converse proposition may not be true: just because growth rate and heterozygosity are correlated, we cannot conclude that the population is at a multilocus equilibrium maintained by selection. An alternative to selectionist explanations is that of inbreeding depression, in which growth rate decreases with inbreeding and the number of homozygous genes is simply a reflection of the inbreeding coefficient. Hence, the degree of inbreeding depression would be inversely correlated with the number of heterozygous loci. This model gains support from the observation that most

studies showing a positive relation between heterozygosity and growth rate also have a deficiency in the number of heterozygotes, which is consistent with a Wahlund effect (Chapter 6). When the pitch pine data (Ledig et al. 1983) were analyzed using a model that assumes multiplicative overdominance (Smouse 1986), more variance could be explained than by using a simple correlation of individual growth rate with individual heterozygosity (Bush et al. 1987). Implicit in this analysis is the assumption that growth rate and fitness are correlated, which is probably acceptable for organisms like clams and pine trees.

SEX RATIO

The evolution of the sex ratio received curiously little attention from early evolutionary biologists, and on what must have been a bad day for Darwin, he simply put the problem aside, writing that "the whole problem is so intricate that it is safer to leave its solution for the future" (*Descent of Man*, p. 399). Fisher (1958, pp. 158–159) made a verbal argument, independent of any genetic mechanism, that is simple but remarkably robust. Fisher's principle is that the sex ratio at reproductive age evolves to the point where "parental expenditure" is the same for both sexes of offspring. In the case of equal cost of producing males and females, this would result in a 1:1 sex ratio. With equal cost, if there were an excess of females in the population, then parents that produce an excess of males will have their genes represented to a larger degree in subsequent generations; and if the male-skewed sex ratio is genetically transmitted, then the population sex ratio must shift back from the female excess. Conversely, if the sex ratio had male excess, then parents producing excess females would be favored. Fisher explained his concept of parental expenditure by means of an example. In species with a long period of parental care and a higher rate of male mortality (such as humans), the expected cost of a male infant is lower than that of a female infant. Thus, for the same cost to the parents, more male infants can be reared than female infants, and hence parents that produce an excess of males will be at an advantage. To say the same thing in another way, if infants were produced in a 1:1 sex ratio, but there was greater male mortality, then there must be an excess of females at reproductive age. Since both sexes must contribute equally to the offspring, whatever the sex ratio, parents that produce extra males would have a greater portion of their genes transmitted to future generations, including their excess male-producing genes. Crow and Kimura (1970, pp. 288–293) present a nongenetic model in support of Fisher's argument, but the precise genetic mechanism determining the sex ratio is critical in models of sex ratio evolution.

An important feature in Fisher's argument is that both sexes contribute equally to their offspring, including genes that determine sex. Hamilton (1967) pointed out that Fisher's argument fails in situations such as haplodiploidy, in which females are diploid and males haploid (e.g., ants, wasps, and bees). Several models incorporating an explicit genetic mechanism of sex determination have been analyzed, but the one that yields both Fisher's principle and fitness maximization is due to Uyenoyama and Bengtsson (1979). They supposed that there are three female genotypes *AA*, *Aa*, and *aa* with respective male offspring frequencies α, β, and γ, and brood sizes σ_1, σ_2, and σ_3. Because males are passive carriers of the *A* genes and do not themselves control sex ratio, this model is valid for autosomal, X-linked, and haplodiploid sex determination.

Let the frequencies of *AA*, *Aa*, and *aa* females be f_1, f_2, and f_3; and the frequencies in males be m_1, m_2, and m_3, where $\Sigma f_i = \Sigma m_i = 1$. Since the only sex ratio control is through females, the fraction of male offspring is

$$M = \frac{f_1\sigma_1\alpha + f_2\sigma_2\beta + f_3\sigma_3\gamma}{f_1\sigma_1 + f_2\sigma_2 + f_3\sigma_3} \tag{9.31}$$

and the fraction of female offspring is $F = 1 - M$. In terms of genotype frequencies, the model can have three internal equilibria, one of which is symmetric with $m_i = f_i$, and two asymmetric equilibria with different genotype frequencies in the two sexes. The following is a necessary and sufficient condition for the existence of the symmetric equilibrium:

$$\sigma_1\sigma_2(\alpha - \beta) + \sigma_2\sigma_3(\beta - \gamma) + \sigma_1\sigma_3(\gamma - \alpha) = 0 \tag{9.32}$$

The equilibrium allele frequency in males for the symmetric equilibrium has the familiar form $p = (\sigma_3 - \sigma_2)/(\sigma_1 - 2\sigma_2 + \sigma_3)$, which is like the two-allele equilibrium with viability selection. The critical results come from the asymmetric equilibria ($m_i \neq f_i$). Uyenoyama and Bengtsson (1979) prove that the mean fitness is maximized at the stable asymmetric equilibrium, and that this equilibrium corresponds to Fisher's equality of parental expenditure argument. To see the latter, suppose that the ratio of parental expenditure on female offspring compared with male offspring is ϕ:1. Females of genotype *AA* therefore have a mean expenditure per offspring of $\alpha + \phi(1 - \alpha)$, and the brood size σ_1 must be proportional to the reciprocal of this. At equilibrium the females of all three genotypes must have the same total parental expenditure, so

$$\sigma_1[\alpha + \phi(1 - \alpha)] = \sigma_2[\beta + \phi(1 - \beta)] = \sigma_3[\gamma + \phi(1 - \gamma)] \tag{9.33}$$

Although it is not immediately obvious, Equation 9.33 is equivalent to Equation 9.32, and at equilibrium the sex ratio of males:females is precisely ϕ. The stability of a 1:1 sex ratio in one-locus sex determination models was

also shown in the situation where sex is determined by the genotype of the individual (Eshel 1975), and in the case where there are multiple alleles at the sex-determining locus (Eshel and Feldman 1982; Karlin and Lessard 1983). These models give mathematical and genetic rigor to a verbal argument that Fisher made 60 years previously.

How sex ratio depends on modifier genes was studied by Thomson and Feldman (1975) using a model of the *Sex Ratio* (*SR*) X chromosome in *Drosophila pseudoobscura*. Suppose two X chromosomes are segregating in a population, with XY males producing equal numbers of male and female offspring, and X*Y males producing $(1/2)f(1 + m)$ daughters and $(1/2)f(1 - m)$ sons. The parameter m measures the degree of meiotic drive of the sex chromosomes, with $m = 0$ being Mendelian segregation. (Segregation in females is Mendelian in all genotypes.) The parameter f specifies the fertility of X*Y males relative to XY males, so $f = 1$ means equal fertility. The model has three genotypes of females (XX, X*X, and X*X*) and two genotypes of males (XY and X*Y), and Thomson and Feldman (1975) considered the general case where the genotypes have viabilities $a:h:b$ and $c:d$ respectively. The analysis gives a recurrence system in the five genotype frequencies. Any interior equilibrium can be expressed in terms of the ratio of frequencies of X*Y to XY males (r_m) and the ratio of frequencies of X*X to XX females (r_f). These are

$$r_m = (d/c)r_f \tag{9.34a}$$

and

$$r_f = \frac{h[c + f(1 + m)d] - 2ac}{h[c + f(1 + m)d] - 2bdf(1 + m)} \tag{9.34b}$$

This equilibrium exists and is stable if both the numerator and denominator of r_f are greater than zero. With equal viabilities, these conditions become $f(1 + m) > 1$ and $f(1 + m) < 1$, and the clear contradiction proves that no such equilibrium exists. The case $f(1 + m) = 1$, which might imply no fertility effect and Mendelian segregation, is *neutrally* stable, so that any newly introduced X chromosome favored by meiotic drive will increase in frequency provided $f(1 + m) > 1$. This means that meiotic drive can give an X chromosome an overall advantage even if it is deleterious to fertility. Moreover, if another X chromosome (X**) with fertility and segregation parameters f_2 and m_2 came along, then X** will invade if

$$f(1 + m) < f_2(1 + m_2) \tag{9.35}$$

In words, the second chromosome invades if it distorts sex ratio even further. The unexpected result is that Mendelian segregation of sex chromosomes is very unstable in maintaining a 1:1 sex ratio because it is susceptible to

invasion by meiotic drive. However, Wu (1983a) found some conditions of fitnesses favoring autosomal modifiers that suppress distortion of *Sex Ratio* X chromosomes.

In the absence of viability or fertility effects, chromosomes like the *SR* X chromosome in *Drosophila pseudoobscura* would be expected to increase in frequency to fixation, driving the population extinct. However, many populations of *D. pseudoobscura* and related species have the *SR* chromosome segregating without going to fixation. Curtsinger and Feldman (1980) addressed the maintenance of this polymorphism by estimating the viability and fertility of the genotypes, and found that there was directional viability selection against the *SR* chromosome, no significant reduction in male fertility, and overdominance in fertility in $X^{SR}X$ females. When incorporated into an extension of Thomson and Feldman's (1975) model allowing multiplicative fecundities, the estimates predict a stable equilibrium, consistent with population experiments. On the other hand, Wu (1983b,c) tested the virility of $X^{SR}Y$ males under different conditions (lower temperatures, nonvirgin females, and allowing remating) and found evidence for a substantial reduction in virility of *SR* males.

An issue intimately related to that of sex ratio is sex determination. The great diversity of mechanisms of sex determination among animals leads to a diversity of strategies and evolutionary implications for the sex ratio (Bull 1983). For example, Hamilton (1967) examined situations that violate Fisher's principle of equality of parental expenditure. He noted that the X−Y mechanism of sex determination is also susceptible to Y chromosome drive giving wide departures from 1:1, just as with X chromosome drive. In surveying organisms that regularly show departures from a 1:1 sex ratio, Hamilton (1967) observed that small arthropods with haplodiploid sex determination, gregarious development of groups of siblings, and regular brother–sister mating often have a large excess of females. This situation results in "local mate competition," and it appears that a female will produce the most grandoffspring if she produces just enough sons to inseminate all her daughters. The strength of local mate competition depends on the number of foundress mothers. As Herre (1987) demonstrated, the theoretically expected tendency towards a greater female bias in broods with a single foundress is nicely borne out in fig wasps.

Sex ratio may be determined as a polygenic threshold trait by the aggregate effects of many genes. True polygenic sex determination is evidently rare, although the swordtail fish *Xiphophorus* may qualify (Kallman 1984). Bull (1983) lists three criteria that must be met to conclude that sex is determined in this fashion: (1) there must be a large between-family variance in sex ratio (i.e., greater than binomial); (2) both males and females must have an influence on progeny sex ratio; and (3) artificial selection on sex ratio should yield

a response. Rice (1986) examined a model in which a population with polygenic sex determination is invaded by a male-determining gene that is dominant and epistatic to the polygenic system, and found conditions in which the male-determining gene replaces the polygenic system.

Some nematodes, as well as turtles, alligators, crocodiles, and other reptiles are among the animals with environmental sex determination, whereby the sex of a developing organism is determined by such factors as temperature (Head et al. 1987). Several types of temperature dependence are found in reptiles (Figure 20). Data on population sex ratio of organisms with envi-

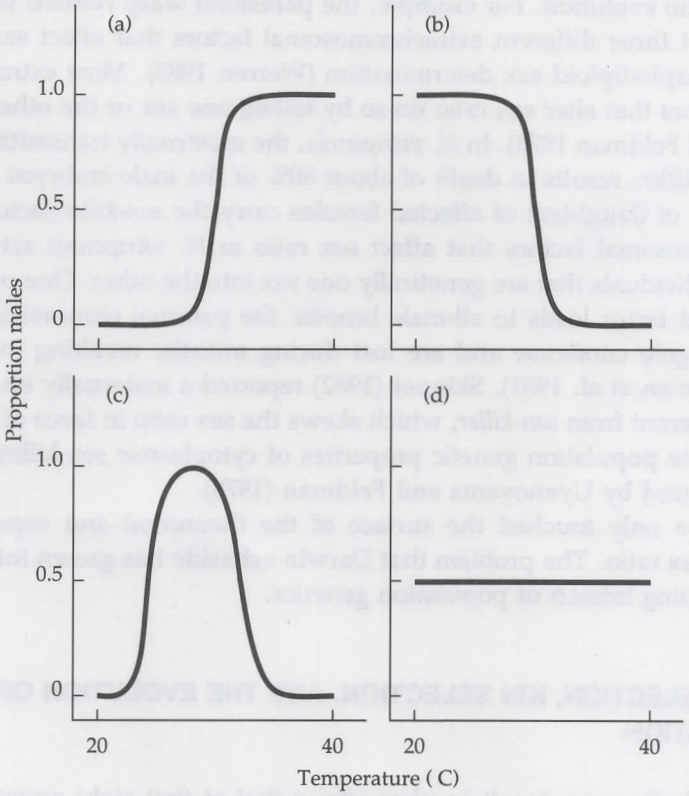


Figure 20. Response of sex ratio to incubation temperature in reptiles. These graphs represent only the approximate form of the four patterns. (a) Females develop at low temperature, males at high temperature, as in two lizards and in alligators. (b) The reverse of (a), with males developing at low temperatures and females at high temperatures (as in most turtles studied). (c) Females at low and high temperatures, males at intermediate temperatures. This pattern may be widespread, as it occurs in snapping turtles and in an Australian crocodile. (d) The hatchling sex ratio of some reptiles is not significantly influenced by incubation temperature. (From Bull 1983.)

ronmental sex determination is scanty, but conditions when environmental sex determination might be adaptive have been studied theoretically (Charnov and Bull 1977; Charnov 1982). Two important conditions are, first, that the environment be patchy, with some patches favoring males and other patches favoring females; and second, that neither parents nor offspring select their patch. When these conditions are violated, genetically determined sex will quickly supplant environmental sex determination (Bull 1983). Plants also have elaborate means of adjusting the relative amounts of energy to male and female functions, including abortion of flowers and fruits (Stephenson 1981).

Extrachromosomal genetic factors can also participate in sex determination and sex ratio evolution. For example, the parasitoid wasp *Nasonia vitripennis* has at least three different extrachromosomal factors that affect sex, in addition to haplodiploid sex determination (Werren 1983). Most extrachromosomal factors that alter sex ratio do so by killing one sex or the other (Uyenoyama and Feldman 1978). In *N. vitripennis*, the maternally transmitted factor called *son-killer*, results in death of about 80% of the male embryos (Skinner 1985); 95% of daughters of affected females carry the *son-killer* factor. Other extrachromosomal factors that affect sex ratio in *N. vitripennis* act by converting individuals that are genetically one sex into the other. One paternally transmitted factor leads to all-male broods: the paternal chromosomes in a diploid zygote condense and are lost during mitosis, resulting in haploid males (Werren et al. 1981). Skinner (1982) reported a maternally transmitted factor, different from *son-killer*, which skews the sex ratio in favor of females. Some of the population genetic properties of cytoplasmic sex-killing factors were analyzed by Uyenoyama and Feldman (1978).

We have only touched the surface of the theoretical and experimental work on sex ratio. The problem that Darwin set aside has grown into a deep and intriguing branch of population genetics.

GROUP SELECTION, KIN SELECTION, AND THE EVOLUTION OF COOPERATION

Natural selection can result in phenomena that at first sight appear to decrease fitness. Animals sometimes behave in a way that makes no sense from the point of view of maximizing individual fitness. Because of the difficulties in measuring behavior, and the great opportunity for behavior to be modified by experience, investigators who adopt the "adaptationist program" and try to find adaptive reasons for behavior must take special pains to remain objective. However, the commonness of group interactions and cooperation

requires an explanation, and some innovative theoretical work has been done to address the problems. As you will see, the theory also poses questions that only careful empirical work can answer.

Group Selection

The term **group selection** has been used loosely to describe almost any kind of selection other than individual selection — that is to say, group selection is often used to denote any kind of genetic change brought about by the differential extinction or proliferation of populations. In this section, it will be convenient to use the term in a more restricted sense to refer to selection that increases the frequencies of genes beneficial to the group as a whole, but detrimental (or at best selectively neutral) to the individuals that carry them. This more restricted definition of group selection we will call **group selection in the strict sense**. (For discussions of the varieties of group selection, see Boorman and Levitt 1980; Levin and Kilmer 1975; Leigh 1977; and Wade 1978, 1985.)

The rationale for group selection in the strict sense is developed in the context of the regulation of population growth. In many species, excessive population growth of a deme can cause extensive damage to the deme's environment; this leads to an inevitable population crash and perhaps even extinction of the deme. Such might be the fate of a predator population that excessively depleted its prey, for example; or the reverse, as in North American deer, whose natural predators have largely been removed by human activity. The deer can increase dramatically in population size, only to be cut back ruthlessly by winter starvation, especially of fawns (for discussion see Smith 1974). Some species, particularly among birds and mammals, have evolved elaborate behavioral mechanisms, such as pecking orders and other social hierarchies or the establishment and defense of territories that, in times of population excess, serve to prevent reproduction of certain members of the population. The basic idea behind group selection in the strict sense is that genes for mechanisms of population regulation are selected because in preventing overpopulation they benefit the entire group, even though the genes are surely detrimental to the fitness of those individuals excluded from reproduction. When there are many demes, there may be genetic variation among demes in the extent of population regulation. Demes with genes that regulate population size at an optimal level will prosper at the expense of demes that overpopulate the environment and become extinct or at the expense of demes that reproduce insufficiently to perpetuate themselves. In this manner the genes for population regulation are spread throughout the species (Wynne-Edwards 1962).

Many population geneticists doubt that group selection in the strict sense used here can occur except perhaps under unusual circumstances. The main problem is in accounting for the evolution of population regulation within demes. How can a gene that causes its bearers to refrain voluntarily from reproduction increase in frequency? Since population self-regulation and other traits beneficial to groups have evolved, skeptics of group selection in the strict sense have searched for alternatives to account for the evolution. These alternatives are based either on individual selection of some kind or on such modes of selection as kin selection (discussed below) that do not seem so anti-Darwinist. One simple alternative is to suppose that the self-regulating mechanisms of populations may not always be selected directly but rather may arise as a correlated response to individual selection for some other trait or complex of traits. An even simpler alternative is an explanation based directly on individual selection. For example, individual birds that establish and defend a territory surely enhance the transmission of their own genes; the genes for territorial behavior will therefore increase in frequency due solely to individual selection. That such behavior is ultimately beneficial to the group is not the essential cause for its evolution but a happy coincidence.

Optimal population size is only one trait that has been hypothesized to evolve according to some mode of group selection. Other traits include rates and patterns of migration, mating, competition, feeding and predation, and especially patterns of behavior of individuals toward other members of the same group. The mode of group selection envisaged, however, is not always group selection in the strict sense.

The case for group selection is best made empirically, and studies with the flour beetle *Tribolium castaneum* have indicated the greatest strength of group selection (Wade and McCauley 1980; McCauley and Wade 1980; Wade 1982). The studies use populations split into demes that are either completely isolated or connected by various levels of migration. The group phenotype that is studied is population productivity or population size. The term "population heritability" has been used to measure the proportion of the observed between-deme phenotypic variance that is heritable in the sense that demes descended from a single parental deme resemble each other (Wade and McCauley 1980; Slatkin 1981b). The hypothesis that there is transmissible phenotypic variation among demes is actually the crux of group selection (Lewontin 1965). In the *Tribolium* studies, the population heritability was found to be significantly different from zero under a variety of conditions, including migration rates up to 12.5% per generation. The parallelism between group selection and covariance among relatives (implicit in the term heritability) was generalized to an analysis of covariance of genic effects in order to partition the effects of selection into within-group and between-

group components (Wade 1985). This partitioning also emphasizes the similarities and differences between selection in heterogeneous environments (hard and soft selection) and group selection. None of these phenomena occur in large panmictic populations, and they all are aspects of the interdeme selection so important in Wright's shifting balance theory (Uyenoyama and Feldman 1980).

Kin Selection and the Evolution of Altruism

Kin selection is a special kind of group selection that has been postulated in attempts to account for the evolution of altruistic behavior. It refers to indirect selection for alleles that occurs through the relatives of carriers of those alleles rather than by direct selection through an enhanced fitness of the carriers themselves. **Altruism** consists of behavioral traits or other attributes that increase the fitness of other individuals at the expense of one's own fitness. Altruistic behavior is exhibited most dramatically by such social insects as termites, ants, and bees, in which certain worker castes exert their labors for the care, protection, and reproduction of the queen and her offspring, but do not reproduce themselves. Other less dramatic examples of altruistic behavior include phenomena such as parental care. As an example of parental care, consider a female deer, who will often flaunt herself before a wolf or other predator when the predator is too near her fawn. The display distracts the predator, who may then pursue the mother; the mother thus places herself at risk in order to protect the fawn, so the parental behavior in this case is altruistic. A dramatic example of altruism is the "adaptive suicide" of pea aphids, which if parasitized by a braconid wasp will drop from the plant when threatened by another predator. This increases their own mortality, but distracts the predator from uninfected neighbors (McAllister and Roitberg 1987).

Haldane (1932) was the first to consider the evolution of altruism. He realized that the magnitude of the risk undertaken should depend on the closeness of the genetic relationship between the altruist and the recipient of the altruism. A central consideration in kin selection is that relatives have genes in common; thus, a gene that causes altruistic behavior can increase in frequency if the increase in the recipient's fitness as a result of altruism is sufficiently large to offset the decrease in the altruist's own fitness. The essentials of the situation can be made clear by considering the case of identical twins. Because identical twins are genetically identical, the reproduction of one's twin is genetically equivalent to reproduction by oneself. Thus it makes no difference if an altruistic individual decreases its own fitness for the sake of an equal increase in fitness of an identical twin; from an evolutionary point of view, it is an even trade because the combined number

of offspring from both twins remains unchanged. By the same token, if an altruistic act decreases the fitness of an individual by an amount less than the increase gained by an identical twin, then the altruism results in a net increase in the combined number of offspring. One would, therefore, expect altruism between identical twins to be favored by natural selection as long as the risk to the altruist is no greater than the benefit to the recipient.

The above reasoning can be made quantitative and extended to other degrees of relationship by referring to Figure 21. In Figure 21a, the circle labeled I represents an individual heterozygous for the *A* allele. (For ease of argument, we suppose that the *A* allele is rare, so that *AA* homozygotes can be neglected, and it is assumed that there is no inbreeding.) The circle labeled R in Figure 21 represents a relative (genotype unknown) of the *Aa* individual, with the degree of relationship between the two individuals measured by a quantity denoted r in the figure. The quantity r is the probability that R carries the *A* allele when we know that I carries that allele. In the absence of inbreeding, r is called the **coefficient of relationship** and equals two times the probability that a randomly chosen gamete from I and a randomly chosen gamete from R carry alleles that are identical by descent (the latter probability is, by definition, the inbreeding coefficient of a hypothetical offspring of I and R).

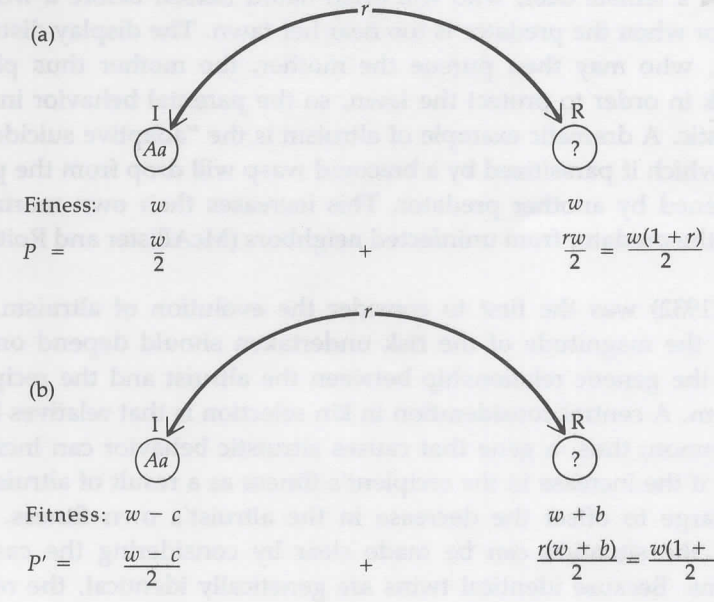


Figure 21. Calculation of inclusive fitness of individual I with genotype *Aa* based on the fitness of I (w), that of a relative (R), and on the coefficient of relationship between I and R (r). (a) No altruism. (b) I is an altruist who increases R's fitness by an amount b at a cost of c units in fitness to himself.

Now, what is the overall number of A alleles contributed to the next generation by the individuals in Figure 21a? The fitness of each individual in the figure is given as w, so the number of A alleles contributed to the next generation, P, is $P = w/2$ (from individual I, the 1/2 entering because only half the offspring receive the A allele) + $rw/2$ (from relative R, recalling that r is the probability that R carries the A allele, given that I does; here again the 1/2 enters because, when R carries A, the probability that it will be included in a particular gamete is 1/2). Such fitness calculations that take into account not only an individual's own fitness but also the fitness of relatives (other than direct descendants) constitute what is often called the **inclusive fitness** of the individual; inclusive fitness therefore represents the total effect of both individual and kin selection.

Figure 21a depicts a situation with no altruistic behavior. Altruism is represented in Figure 21b, where we assume that the A allele leads to an altruistic behavior that reduces I's fitness by an amount c while increasing R's fitness by an amount b. You can think of c as the *c*ost and b as the *b*enefit of the altruism. The same reasoning as before shows that, in this case, the overall number of A alleles contributed to the next generation, denoted P', is

$$P' = (w - c)/2 + [r(w + b)]/2 \qquad \textbf{(9.36)}$$

Altruism will be favored by natural selection if $P' > P$—that is to say, if $(w - c)/2 + [r(w + b)]/2 > w/2 + rw/2$. A little manipulation reduces this inequality to the simple statement

$$c/b < r \qquad \textbf{(9.37)}$$

In other words, a gene for altruism toward a relative will be favored if the ratio of the loss in fitness of the altruist to the gain in fitness of the relative is less than the coefficient of relationship. This relationship, known as Hamilton's Rule (Hamilton 1964), has been widely applied without careful regard for the assumptions that went into its derivation. The need for careful analysis is emphasized by work of Cavalli-Sforza and Feldman (1978) and Charlesworth (1978), who presented models for the evolution of altruism in which altruistic alleles can invade when Hamilton's Rule is violated or altruistic alleles cannot invade even though Hamilton's Rule is satisfied. Cavalli-Sforza and Feldman (1978) eschew the concept of inclusive fitness and show that kin selection can be defined with greater precision as a form of frequency-dependent selection. Hamilton's Rule is valid in the special case when costs and benefits to fitness are additive; when they are nonadditive, the rule fails to predict when an altruistic allele is favored. A weaker version of Hamilton's Rule is that increasing the degree of relatedness increases the likelihood of altruism being favored. However, even this property is violated when groups consist of related individuals and unrelated "intruders" (Matessi and Karlin

1986). The literature on kin selection and the evolution of altruism is growing rapidly, and some additional important papers include Eshel (1977), Matessi and Jayakar (1976), Levitt (1975), Michod (1982), Sherman (1980), Uyenoyama (1984), and Uyenoyama et al. (1981).

F Kin Selection in Haplodiploid Insects

Kin selection is somewhat complicated in species of the insect order Hymenoptera (e.g., ants, wasps, and bees) because of their peculiar mode of reproduction. In ants, bees, and most other hymenopteran insects, females originate in the normal manner by fertilization of a haploid egg by a haploid sperm; males, however, originate by parthenogenetic development of unfertilized haploid eggs (Hartl and Brown 1970; Kerr 1974; Crozier 1977). The peculiar feature of male haploidy is that a female is more closely related to her sisters than she is to her own daughters. An illustration of this point is found in Figure 22, where we make the simplifying assumptions that there is no inbreeding and that females mate with a single male. Alleles that are denoted by the α terms with different subscripts (α_1 and α_3, for example) are assumed not to be identical by descent. The genotype of the female (I) at the top of the pedigree is designated $\alpha_1\alpha_2$, and her daughter's genotype (J) is arbitrarily assumed to be $\alpha_1\alpha_3$. Recall that, in the absence of inbreeding, r is the coefficient of relationship between two individuals; if the

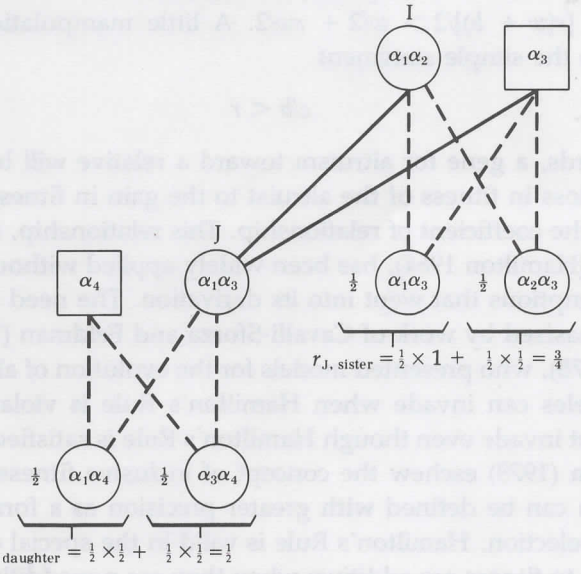


Figure 22. Genetic relationships between organisms in species with haploid males. The coefficient of relationship between the female J and her own daughters is 1/2, whereas the coefficient of relationship between J and her sisters is 3/4. Thus J can bring more of her own alleles into the population by aiding her mother (I) in producing more sisters than she can by reproducing herself, provided that the mother mates with only one male.

two individuals have one allele in common (as for parents and offspring with autosomal inheritance or for mothers and daughters with male haploidy), $r = 1/2$; if the individuals have two alleles in common (as for identical twins), $r = 1$.

The coefficient of relationship between J and her daughters is calculated in Figure 22 as 1/2 [which is the probability that the daughter is genotypically $\alpha_1\alpha_4$] $\times$ 1/2 [which is the coefficient of relationship in this case] + 1/2 [the probability that the daughter is $\alpha_3\alpha_4$] $\times$ 1/2 [the coefficient of relationship in this case]; $r_{J,daughter}$ therefore works out to be 1/2.

As between individual J and her sisters, the coefficient of relationship is calculated in Figure 22 as 1/2 [the probability that the sister is genotypically $\alpha_1\alpha_3$] $\times$ 1 [the coefficient of relationship in this case] + 1/2 [the probability that the sister is $\alpha_2\alpha_3$] $\times$ 1/2 [the coefficient of relationship in this case]; it thus turns out that $r_{J,sister} = 1/2 + 1/4 = 3/4$.

Since the coefficient of relationship between J and her sisters is greater than that between J and her own daughters, it stands to reason that, from an evolutionary point of view, J would be better off to produce a sister than to produce a daughter. But how can J produce a sister? She can do so by promoting further reproduction by her own mother — the queen. The startling fact that hymenopteran females are more closely related to their sisters than to their own offspring may explain why true social behavior has evolved independently several times in the Hymenoptera, although it is rare elsewhere among animals. Moreover, for a hymenopteran male, the probability that his sister carries the same allele as he does is 1/2, whereas the probability that his daughter carries the same allele is 1. Thus hymenopteran males are more closely related to their daughters than they are to their sisters, and a male hymenopteran would consequently be better off to have a daughter of his own than to promote further reproduction by the queen. This kind of reasoning may explain why, when social behavior has evolved in the Hymenoptera, the sterile worker castes that aid the queen are always female. (For more on social insects, see Kerr 1969; Wilson 1971; Hamilton 1972; Trivers and Hare 1976; and Oster et al. 1977.)

Arguments such as those above that attempt to explain the evolution of social behavior are important ingredients in the field of evolutionary sociobiology. **Sociobiology** itself is defined as the systematic study of all social behavior, particularly animal social behavior (Wilson 1975). Evolutionary sociobiology is therefore the study of the mechanisms by which social behavior evolves. In the case of social behavior among the Hymenoptera, considerations of kin selection play a central role in understanding the evolution of sociality. Kin selection of some sort, and perhaps other modes of group selection as well, may play a role in the evolution of social behavior in other organisms, such as primates. It should be noted, however, that it is one thing to provide a theoretical demonstration that social behavior could evolve by means of kin selection (or any other mode of selection), but it is quite another thing to prove that the behavior actually did evolve according to the

theoretical mechanism. The actual manner of evolution of social behavior can be determined only by experiment, observations of organisms in their natural habitats, and comparisons of social behavior among related organisms.

Evolution of Cooperation

In the discussion of kin selection and the evolution of altruism, the altruistic acts were assumed to be harmful to the altruist. What about behavior that does not reduce fitness? Perhaps cooperative behavior is favored even with unrelated individuals. The model in Figure 21b is appropriate for cooperation if the cost c is negative and the benefit b is positive. The condition $P' > P$ is still $c/b < r$, but note, since c is negative, $c/b < 0$. Thus, cooperative behavior is favored for any positive value of r and, indeed, for $r = 0$, which corresponds to I and R being unrelated. An example of cooperative behavior is found in large predators such as wolves, which hunt in packs. Hunting in packs increases the probability of a successful kill, and all animals in the pack are equally benefited. This sort of cooperative behavior is promoted by natural selection, even if the cooperating individuals involved are unrelated.

The evolution of cooperation has been studied using a game known as the Prisoner's Dilemma (Rapoport and Chammah 1965). Two individuals A and B must choose either to cooperate or defect (Figure 23). If they both cooperate, the "payoff" to each is X, whereas if both defect, the payoff is Y, where $Y < X$. The rub comes when one cooperates but the other defects, in which case the defector gets payoff W (the greatest of all payoffs) and the cooperator gets payoff Z (the smallest payoff). Since the payoffs satisfy the inequality $W > X > Y > Z$, the game is a dilemma because if one cooperates and the other defects, the defector gets the greatest gain, while the cooperator gets the least payoff; this makes defection sound pretty attractive, but if both players defect, their payoff is less than if they had both cooperated. Of course, the players could agree to alternately cooperate and defect (which would be a sort of cooperation), but this is also a nonoptimal strategy if $2X > W + Z$.

Axelrod and Hamilton (1981) examined a simple model for the evolution of cooperation using the Prisoner's Dilemma game with repeated encounters. As in real life, genotypes can adopt different strategies depending on previous experience. The winning strategy is one in which you start with cooperation, but if the other fellow defects, you copy his preceding play on the next encounter. This strategy has the name TIT FOR TAT. Because TIT FOR TAT starts out cooperating, a strategy of pure defection, called ALL DEFECT, will beat TIT FOR TAT unless the chance of a repeat encounter is sufficiently high. If the chance for a repeat encounter is greater than $(W - X)/(W - Y)$, then TIT FOR TAT wins. Axelrod and Hamilton's (1981) model assumes a

		B: Cooperate	B: Defect
A	Cooperate	X, X	Z, W
A	Defect	W, Z	Y, Y

Figure 23. The payoff matrix for the Prisoner's Dilemma. The players A and B may choose either to cooperate or to defect. In each cell of the matrix is the payoff to player A and B respectively. If both cooperate, each gets payoff X, and if both defect, each gets Y (which is less than X.) If A cooperates, but B defects, A gets payoff Z and B gets payoff W, where $W > X > Y > Z$.

haploid genetic transmission of strategies, and they claimed that the TIT FOR TAT strategy is an **evolutionarily stable strategy (ESS)**. An ESS is one that resists invasion by other strategies (Maynard Smith 1982). The claim that TIT FOR TAT is an ESS has been contested by Boyd and Lorberbaum (1987), who show that the TIT FOR TAT strategy can be beaten if the population is simultaneously invaded by two other strategies. They go on to prove that no pure strategy can be an ESS. More complexities are being built into models for the evolution of cooperation by making the payoff matrix depend on the genetic constitution of the population (Peck and Feldman 1986), and by considering the strategy of ceasing to play (Feldman and Thomas 1987). Stable polymorphisms are possible in both of these situations.

COEVOLUTION

In assuming constant fitnesses, many of the models we have examined ignore one of the most wonderful aspects of biological evolution: that interactions among species shape their evolutionary fates. The term **coevolution** was coined by Ehrlich and Raven (1964) to describe the evolutionary consequences of interactions between plants and herbivorous insects. In these interactions, plants evolve a variety of chemical defenses to avoid insect infestation, and the insects counter with mechanisms to detoxify or otherwise resist the plants' chemical warfare. A recurring theme in the evolution of chemical defense is that a good strategy is to produce a chemical that is similar or identical to a vitally important substance in the creature being defended

against. This strategy makes it difficult for the herbivore or parasite to circumvent the chemical. A striking example of this strategy is found in a Malaysian plant of the genus *Cyperus*, which synthesizes the insect juvenile hormone JH III (Toong et al. 1988). The JH III produced by the plant is biologically active, since 90% of grasshoppers reared on the plant have defective wings and underdeveloped ovaries. Juvenile hormone is a sesquiterpenoid, and the synthetic pathway evolved in this plant is a remarkable example of the chemical sophistication of biochemical coevolution. In the following, only a few types of interspecific interactions will be examined; an excellent guide to the literature is Futuyma and Slatkin (1983).

Mimicry

Many insects exhibit a kind of coloration called **warning coloration**, involving bright, often gaudy, highly conspicuous colors and patterns. These insects are usually extremely distasteful to vertebrate predators such as birds and lizards that hunt for prey primarily by sight. The warning colors flaunt the presence of the insects, the predators having learned from prior unpleasant experience to leave them alone. (The distastefulness of the insects often comes from chemicals in the plants they eat as larvae.)

In the mid-nineteenth century, the English naturalist Henry W. Bates noticed a curiosity of butterfly coloration while making collections in Brazil; certain atypical species, whose near relatives were cryptically colored, exhibited the bright warning colors and patterns of quite unrelated species living in the same habitat. Bates reasoned that the atypical species were palatable **mimics**, using their coloration to disguise themselves and masquerade as members of the warningly colored, unpalatable species (the **models**), thereby deceiving their predators. This phenomenon, in which a palatable species evolves the warning coloration of an unpalatable one in order to escape predation, has come to be known as **Batesian mimicry** (for reviews see Sheppard 1959; Turner 1977).

Batesian mimicry is only one of several kinds of mimicry. Certain insects mimic twigs or leaves, for example. Plants are capable of mimicry too: besides the synthesis of juvenile hormone, some plants produce small protuberances on their leaves that resemble butterfly eggs, convincing the female butterfly that the leaves are already occupied; and some orchids have flowers that mimic the abdomen of female bees in order to attract male bees for pollination. An important kind of mimicry found especially in tropical butterflies is called **Müllerian mimicry** after Fritz Müller, a nineteenth century German zoologist (see Turner et al. 1979). In Müllerian mimicry, as in Batesian mimicry, two or more unrelated species will exhibit strikingly similar warning colors and patterns; in contrast to Batesian mimicry, all the species involved

in Müllerian mimicry are unpalatable, their mutual mimicry serving to protect them collectively from predation inasmuch as a predator need learn to avoid only a single pattern of coloration instead of many. Fewer individuals, then, are lost to the predator's learning experience.

An important premise behind the theories of Batesian and Müllerian mimicry is that vertebrate predators do indeed learn to avoid noxious prey. That this is the case is abundantly clear from the data shown in Figure 24, which are the results of an experiment in which artificial cylindrical "caterpillars" composed of lard and flour and either red or blue food coloring were offered as bait to wild birds, principally blackbirds, song thrushes, house sparrows, dunnocks, and robins. As shown in Figure 24, the birds initially ate (or pecked) the red and blue "caterpillars" equally. When the red baits were made noxious by the addition of quinine sulfate, however, the birds quickly learned to avoid them; by the end of the sixth day only about 2% of the baits eaten or pecked were red. Moreover, when birds that had learned to avoid

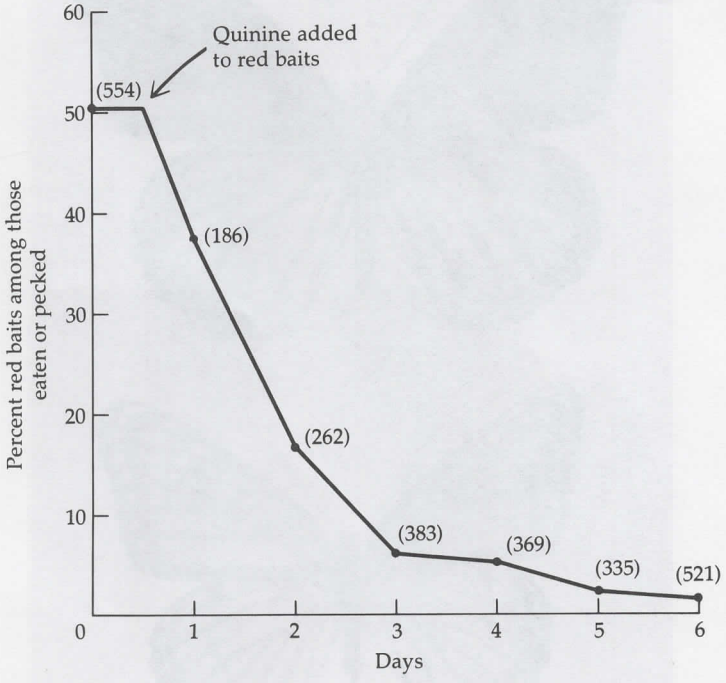


Figure 24. Wild birds quickly learn to avoid red baits that have been rendered distasteful to them by the addition of quinine. The ordinate gives the percentage of red (as opposed to blue) baits eaten or pecked, and is plotted against days. Numbers by the points are total number of baits eaten or pecked on that day. (Data from Ford 1971.)

quinine-flavored red baits were presented with unflavored bait that was half red and half blue, they carefully ate away the blue half, but left the (to them unknown) equally tasty red half behind.

For Batesian mimicry to occur, of course, the predator must not only learn to avoid the unpalatable model, it must also avoid the palatable but deceptively colored mimic. Whether predators actually are duped by mimics has been examined in a famous case of Batesian mimicry involving two large orange and black butterflies found in the eastern United States, the monarch butterfly (*Danaus plexippus*), which is the model, and the viceroy butterfly (*Limenitis archippus*), which is the mimic (Figure 25). In these experiments, butterflies were fed to caged Florida scrub jays (Brower 1958). Birds fed the monarch (the model) reacted violently to its bad taste and quickly learned not to eat it; thereafter they also avoided the viceroy (the mimic). On the

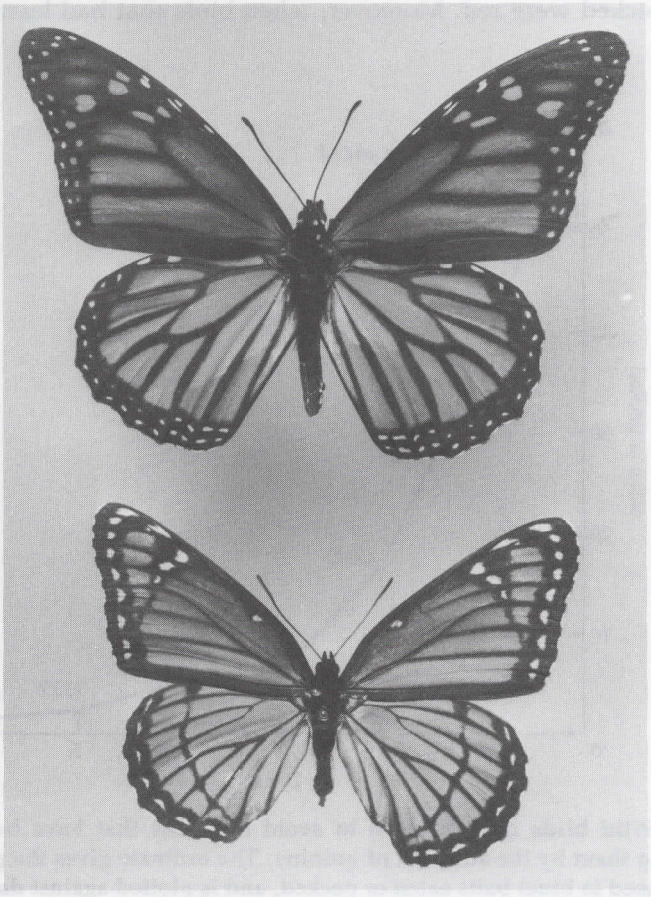


Figure 25. The monarch butterfly *Danaus plexippus* (the model) and the viceroy butterfly *Limenitis archippus*, its Batesian mimic. (Courtesy of T. Eisner.)

other hand, birds never fed the monarch would eat the viceroy. (Such birds did not feed on the viceroy very happily, though; when given a choice between viceroys and unrelated, highly palatable butterflies, the birds spurned the viceroys. This shows that "palatability" is really a matter of taste and that in the monarch–viceroy example, the distinction between Batesian and Müllerian mimicry is slightly blurred.)

In any case, the sort of evidence just described shows that predators do behave as Bates had originally surmised. There is, however, another important test of the theory of Batesian mimicry: if mimics really do mimic a model because of the selective advantage it confers, then butterflies that live in habitats beyond the geographical range of the model ought not to be mimetic. A splendid illustration of this expected geographical correlation is provided by the highly distasteful butterfly model, *Battus philenor*, and its palatable Batesian mimic, *Limenitis arthemis*. The map in Figure 26 shows the essential

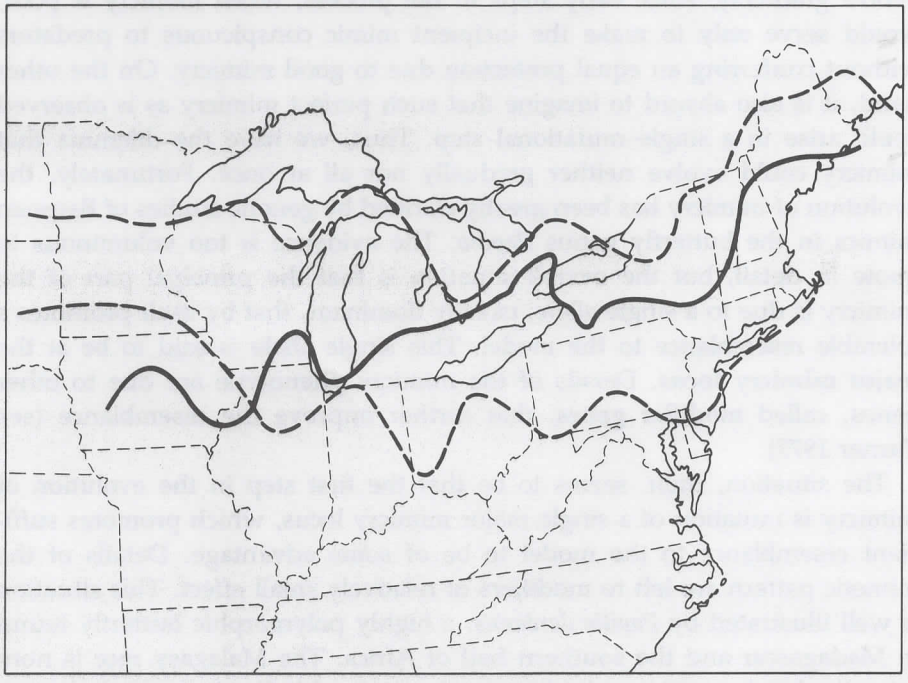


Figure 26. The northern limit of the extremely distasteful butterfly model *Battus philenor* (solid line) in the United States coincides with Batesian mimicry by its mimic *Limenitis arthemis*. Below the lower dashed line *L. arthemis* is an excellent mimic; above the upper dashed line there is no mimicry of *B. philenor*. In the intermediate zone between the dashed lines, coinciding with the geographical limit of *B. philenor*, the quality of mimicry is extremely variable due to hybridization between mimetic and nonmimetic forms. (After Platt and Brower 1968.)

situation: the solid line depicts the approximate northern limit of the model, while the dashed lines pertain to *L. arthemis*. Below the lower dashed line (at about 42°N latitude), *L. arthemis* mimics *B. philenor* in having extreme blue-green iridescence of the dorsal hind wings; above the upper dashed line (at about 43°N latitude), *L. arthemis* is nonmimetic, lacking the iridescence but having white bands and extensive red-orange spotting on the wings. Between these two limits the mimetic and nonmimetic forms hybridize, and the butterflies found there are extremely variable, some being rather good mimics while others are very poor mimics. The important point here is that the excellent mimicry found in *L. arthemis* is strikingly correlated with the presence of its model.

Evolution of mimicry

Mimicry is a complex adaptation involving many aspects of morphology and behavior (behavior because mimics often imitate such behaviors of models as flight or resting postures). Such an adaptation could not be expected to evolve gradually, since early steps in the process, when mimicry is poor, would serve only to make the incipient mimic conspicuous to predators without conferring an equal protection due to good mimicry. On the other hand, it is also absurd to imagine that such perfect mimicry as is observed could arise in a single mutational step. Thus, we have the dilemma that mimicry could evolve neither gradually nor all at once. Fortunately, the evolution of mimicry has been greatly clarified by genetic studies of Batesian mimics in the butterfly genus *Papilio*. The evidence is too voluminous to quote in detail, but the overall situation is that the principal part of the mimicry is due to a single allele, usually dominant, that by itself promotes a tolerable resemblance to the model. This single allele is said to be at the **major mimicry locus**. Details of the mimicry phenotype are due to other genes, called **modifier genes**, that further improve the resemblance (see Turner 1977).

The situation, then, seems to be that the first step in the evolution of mimicry is mutation of a single major mimicry locus, which promotes sufficient resemblance to the model to be of some advantage. Details of the mimetic pattern are left to modifiers of relatively small effect. This situation is well illustrated by *Papilio dardanus*, a highly polymorphic butterfly found in Madagascar and the southern half of Africa. The Malagasy race is nonmimetic, but in continental Africa many geographically distinct races occur that are Batesian mimics of different models. When various African mimics are crossed to the Malagasy race, the distinctive details of the mimicry are lost in the hybrids, but the major genes, which are dominant and apparently allelic, are expressed and produce almost identical patterns. (Interestingly, only the females in this species are mimetic.) Butterflies make great use of

vision during courtship, and females choose their mates based largely on color pattern; mimicry in males does not evolve because a mimetic pattern in males too greatly diminishes the chance of being chosen for mating.)

A test of the degree of resemblance necessary to confer an advantage can often be made quite directly. In Trinidad, for instance, wing margins of male *Hyalophora promethea* moths were painted black to "mimic" the female of the "model" *Parides anchises*. The moths were released along with unpainted control moths and recaptured, and from the recapture rates a selective advantage due to the artificial mimicry could be calculated. In one such experiment, the selective advantage was around 50%, although the estimate is an approximate one and the advantage depends markedly on the number of mimics relative to the number of models (Brower et al. 1967). Indeed, with too many mimics relative to the number of models, the predators may taste a few and learn that animals with warning coloration are largely palatable; this consideration sets an upper limit to the number of mimics that can be maintained without a breakdown in protection due to mimicry (Charlesworth and Charlesworth 1975).

Modifiers and supergenes

Two types of modifiers must be distinguished: **specific modifiers** are modifiers expressed only in the presence of the major mimicry gene and improve the resemblance to the model; **nonspecific modifiers** are expressed in all genotypes, and they lead to improved mimicry (and hence increased fitness) in the presence of the major gene but merely promote conspicuousness (and hence decreased fitness) in the absence of the major gene. The distinction between specific modifiers and nonspecific modifiers is important because specific modifiers in natural populations typically exhibit no genetic linkage with the major gene, where nonspecific modifiers are usually tightly linked. Moreover, Charlesworth and Charlesworth (1976a,b) have shown that this relationship between type of modifier and genetic linkage is exactly what is expected on theoretical grounds when the major mimicry gene is polymorphic.

The action of specific modifiers can be illustrated by comparison of the African *Papilio dardanus* with a relative in the Far East, *P. polytes*. In both species, populations that are polymorphic for major mimicry genes also contain specific modifiers that alter the lengths of the tails on the hindwings of the mimetic forms. In *P. dardanus*, the modifiers shorten the tails to improve resemblance to the model; in *P. polytes*, the modifiers lengthen the tails. These modifiers of tail length are unlinked to the major mimicry locus, and the occurrence of shortening or lengthening modifiers is limited to populations in which the mimic exhibits either shorter or longer tails, respectively. That is, modification of tail length in the direction of the model

does occur when mimics in either *P. dardanus* or *P. polytes* are crossed to other members of the same population. When outcrossed to other populations in these species that lack the appropriate mimetic forms, modification of tail length does not occur because the nonmimetic populations do not contain the appropriate modifiers.

Specific modifiers need not be linked to the major mimicry locus because they are expressed only in mimetic forms; in nonmimetic genotypes, they cause no decrease in fitness. For nonspecific modifiers, however, which are expressed in all genotypes and do cause a reduction in fitness in nonmimetic forms, the situation is quite different. Unless the major mimicry gene is fixed or nearly fixed in the population, nonspecific modifiers that are unlinked to the major locus will tend to be eliminated by natural selection. Only if the nonspecific modifiers are tightly linked to the polymorphic mimicry locus will the modifiers increase in frequency. Polymorphism for a major mimicry gene therefore creates a sort of evolutionary sieve, which sifts through mutations of nonspecific modifier loci as they occur and preserves only those that are tightly linked. In time, there evolves a **supergene** consisting of the major mimicry locus and around it a cluster of tightly linked nonspecific modifiers. Within such a supergene, recombination is rare and linkage disequilibrium is extreme; so the block of loci causing mimicry is usually inherited as a single genetic unit, which in genetic crosses creates the illusion that nearly perfect mimicry is due to a single locus (Charlesworth and Charlesworth 1976a).

Mimicry supergenes have been well analyzed in the Oriental species *Papilio memnon*, where rare recombinants found in natural populations permit the genetic structure of the supergene to be inferred. Six loci have been identified, all within a recombination distance of 0.6% (see Turner 1977). The component loci in the supergene, given in linear order, control the presence or absence of tails, blue marks on the hindwings, white marks on the hindwings, color pattern of the forewing, color of the basal triangle on the forewing, and color of the abdomen. The situation in *P. memnon* is by no means unique, as the multiple mimetic forms in *P. dardanus* and *P. polytes* also seem to be controlled by supergenes.

In summary, Batesian mimicry is a complex adaptation involving many aspects of morphology and behavior. It is quite clear that the adaptation evolves in two major stages: first, the selection of a major mutation that provides a sudden, often crude, but satisfactory sort of mimicry; and second, the selection of specific and nonspecific modifiers that in small increments progressively improve and finally perfect the resemblance. What is not so clear — and this may ultimately be the most interesting question raised by studies of Batesian mimicry — is how many other major adaptations evolve by means of an analogous two-step process.

Industrial Melanism in Moths

Industrial melanism refers to the evolution of black (melanic) color patterns in certain species of moths that accompanied progressive pollution of the environment by coal soot during the industrial revolution. (The various color forms of the moths are known as **morphs**.) Industrial melanism occurs in England, West Germany, eastern Europe, the United States, and in other heavily industrialized areas. The species involved in industrial melanism are large moths that fly by night and rest in a sort of cataleptic state by day, often on the trunks of trees, using their cryptic black-and-white mottled color pattern for concealment from such visually cued predators as hedge sparrows, redstarts, and robins (Figure 27). Of nearly 800 species of large moths in the British Isles, where industrial melanism has been most intensively studied, about 100 species are industrial melanics (Kettlewell 1973). The best known of these are the peppered moth *Biston betularia* and the scalloped hazel moth *Gonodontis bidentala*.

Industrial melanism evolves extremely rapidly, which is indicative of strong selection. Although melanic *Biston betularia* were first noted in Manchester, England in 1848, their frequency was low (a reasonable guess is 1%), but by 1898 the frequency of melanics in the same area was about 95%. Since *B. betularia* has nonoverlapping generations with a single generation per year, and since the melanic phenotype is basically due to a single dominant allele

Figure 27. Melanic and nonmelanic moths (*Biston betularia*), showing camouflage of light moths on light background and dark moths on dark background. (Photographs by H. B. D. Kettlewell.)

(some minor complications in the genetics are discussed below), the selection coefficient required to account for the observed increase in melanic frequency is on the order of 0.20 (Haldane 1956). Comparable increases in the frequency of melanics as an accompaniment to industrialization have been noted elsewhere. In southeastern Michigan, for example, the frequency of melanics in the related moth *Biston cognataria* increased from 1% to over 80% in about 30 years, again suggesting a selection coefficient of around 0.20.

It has been experimentally confirmed that differential predation of melanic and nonmelanic morphs by birds is an important source of selection. First, the color morphs preferentially rest on backgrounds that make them least conspicuous. (The effect of this behavior can be seen in Figure 27.) For example, when melanic and nonmelanic morphs are released on a tree trunk that is half light and half dark because the dark half has been artificially stripped of lichens and sprayed with soot, the melanic moths tend to rest on the dark surface and the nonmelanics on the light surface (Figure 28a). Second, when marked melanic and nonmelanic moths are released and later recaptured, the proportion of melanics recovered is about twice as great as that of nonmelanics if the release was in a polluted area but only about one-third that of nonmelanics if the release was in a pollution-free area. Third, direct observations of predation by redstarts in light and dark areas showed that in dark areas proportionally more nonmelanic moths were eaten, and

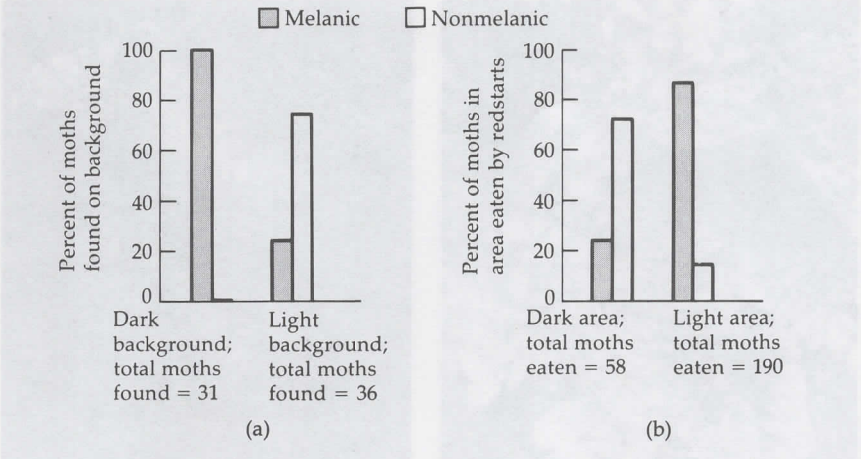


Figure 28. (a) Released moths tend to settle on backgrounds where they are least conspicuous; in this case, the choice was between the light half of a tree trunk and a half that had been artificially "polluted" by stripping it of lichens and spraying it with soot. (b) Birds (in this case redstarts) preferentially eat moths that have settled on backgrounds that make the moth conspicuous. (Data from Kettlewell 1973.)

in light areas proportionally more melanics were eaten — though the moths were released in equal numbers in both areas.

The story that emerged, then, was this: at the beginning of the industrial revolution, when coal was first burned on a large scale, melanic moths were rare because they were heavily preyed upon by birds due to their conspicuousness on light-colored, lichen-covered trees. As soot and other pollutants belched forth from smokestacks and settled over the countryside, first the lichens (which are very sensitive to air pollution) died, then the trees began to turn black. In such blackened areas, the melanic morph was no longer so conspicuous; indeed, the nonmelanic moth, needing light-colored areas for background camouflage, was now at a selective disadvantage. The frequency of the dominant mutation leading to melanic coloration thus began to increase, reaching, as has been noted, about 95% by 1898.

In the 1950s, air pollution became a matter of public concern because of its harmful effects on human health. In England, following the Beaver report of 1954, certain "smokeless zones" were established. In the smokeless zones, rains began to wash and clean the trees. The bark lightened, and lichens began to reappear on tree trunks. Suddenly the melanic form was again at a selective disadvantage, and its frequency began to decline — in one area from 94.8% in 1961 to 91.1% in 1964 to 89.5% in 1974 (Bishop and Cook 1975). Most recently the selection coefficient against the melanics in the smokeless zones has been estimated at 0.12 ± 0.02 (Cook et al. 1986).

A number of complications in this otherwise simple story have been uncovered. First, there are genetic complications. While the principal cause of melanism in *Biston betularia* is a single dominant allele, in certain areas moths neither light nor dark but of intermediate shades of grey are found; these seem to be due to alleles at the same locus, but other loci may also be involved (Lees and Creed 1977). Moreover, the melanic specimens of *B. betularia* collected in the last century are considerably lighter in color than those found today, as if in the course of evolution genetic modifiers leading to greater dominance of the melanic allele have been selected. There are also complications in the selective forces. Although bird predation is undoubtedly an important source of selection, there seem to be differences in viability (and perhaps also fertility) between melanic and nonmelanic morphs that cannot be accounted for by predation. Considering the intensity of selection involved, it is surprising that the allele for melanism never became fixed, suggesting that when the melanic morph is at a high frequency some unidentified form of selection comes into play that tends to maintain polymorphism. Yet another complication is in the mating system, as evidence has been found that the color morphs may undergo a degree of disassortative mating — that is, melanics sometimes mate preferentially with nonmelanics.

Despite these complications, industrial melanism is perhaps the most

widely known case of evolution in animals, demonstrating an immediate and dramatic response to changed environmental conditions. The example also illustrates how genetic differentiation between populations can be impeded by migration. This aspect of the situation is depicted in Figure 29. The peppered moth, *Biston betularia*, is a species that exists in low population density, and individuals must fly relatively long distances to find a mate; as seen in Figure 29a, there is little local differentiation of populations with respect to frequency of melanics: where the frequency of melanics is high it is uniformly high in all populations, and where the frequency is low, it is almost uniformly low. By contrast, the scalloped hazel moth, *Gonodontis bidentata*, exists in much higher densities and migration is less extensive; in this case, there is substantial local differentiation of populations with respect to frequency of melanics (Figure 29b).

The Evolution of Antibiotic Resistance in Bacteria

Plasmids are small, usually circular molecules of DNA that exist inside bacterial cells as self-replicating genetic elements not attached to the bacterial chromosome (see Wolstenholme and O'Connor 1969; Clowes 1972; Meynell 1972; Campbell 1969). Plasmids are heritable agents, transmitted from generation to generation as bacterial cells divide. Many plasmids are also infectious agents, transferable from one bacterial cell to another, even occasionally to cells of different bacterial species. Transfer is mediated by tubelike projections from the cell surface called **pili** (Figure 30), formation of which is often controlled by genes in the plasmid itself. (Bacteria have many different types of pili, only some of which are associated with the transfer of plasmids.)

Plasmids play a central role in the evolution of antibiotic resistance in bacteria, a conclusion that emerged in the late 1950s with the discovery of multiple drug-resistant bacteria. The first such strain was isolated in 1955 from a Japanese woman who developed a stubborn case of dysentery after returning from vacation in Hong Kong (Watanabe 1967). The cause was a typical dysentery bacillus of the genus *Shigella*, but in one way this *Shigella* was unique—it was resistant to four drugs: sulfonamide, streptomycin, chloramphenicol, and tetracycline. Moreover, this strain of *Shigella* was able to transfer its drug resistance to such other intestinal bacteria as *Escherichia coli*.

The troublesome *Shigella* turned out to harbor what has since become known as a **resistance transfer factor**—a plasmid that carries genes conferring resistance to one or more antibiotics as well as genes enabling its own transfer to other bacteria. Since 1955 many other resistance transfer factors have been found in a wide variety of pathogenic and nonpathogenic bacteria. These factors confer resistance to various combinations of drugs (including

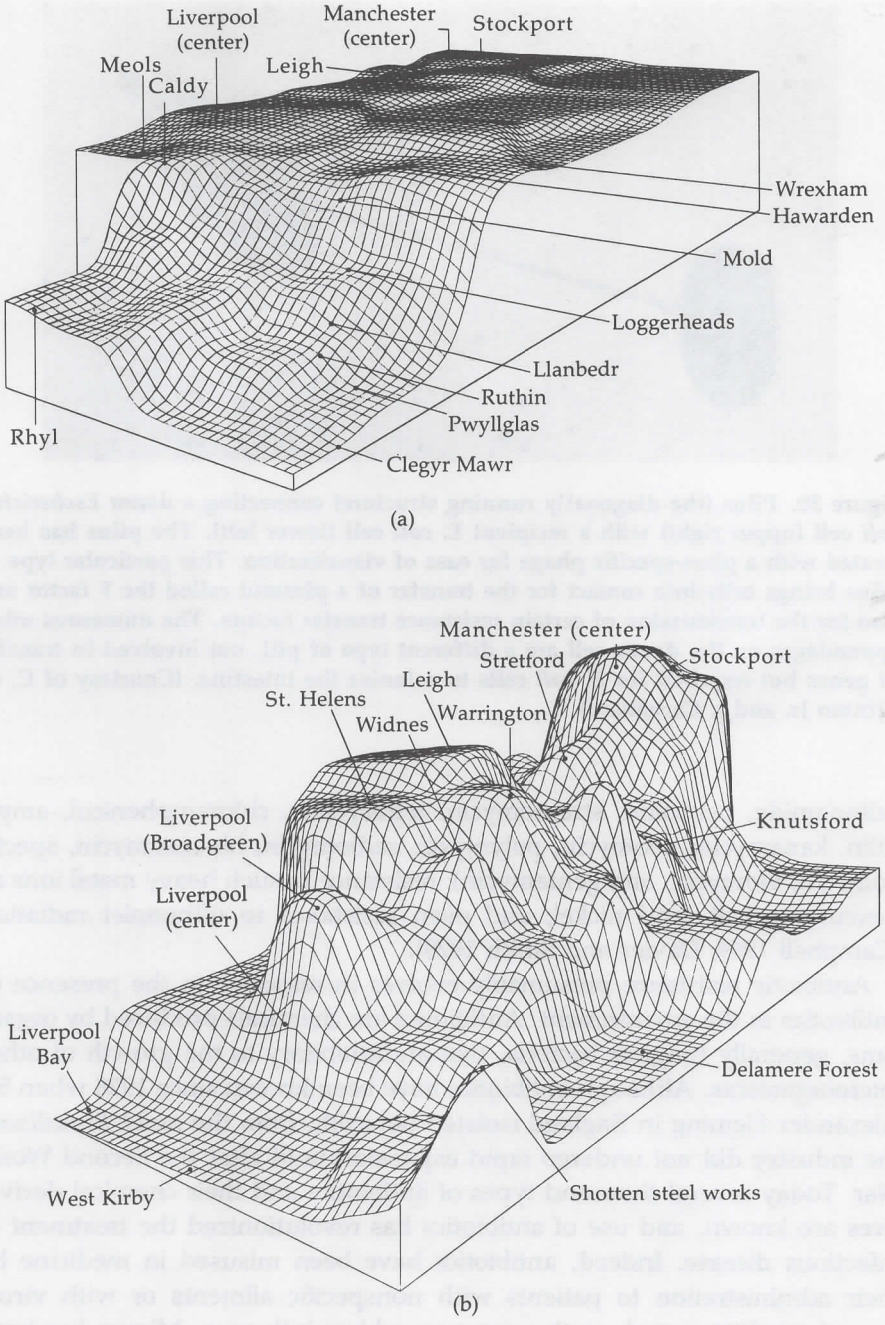


Figure 29. (a) Distribution of melanic individuals of *Biston betularia* over an area including Liverpool and Manchester, as viewed from rural Wales. The vertical axis in these perspective views is the frequency of the melanic morph, while the other two axes are geographic location. (b) Distribution of melanic individuals of *Gonodontis bidentata* over a smaller area than in (a), but viewed from the same perspective. (From Bishop and Cook 1975.)

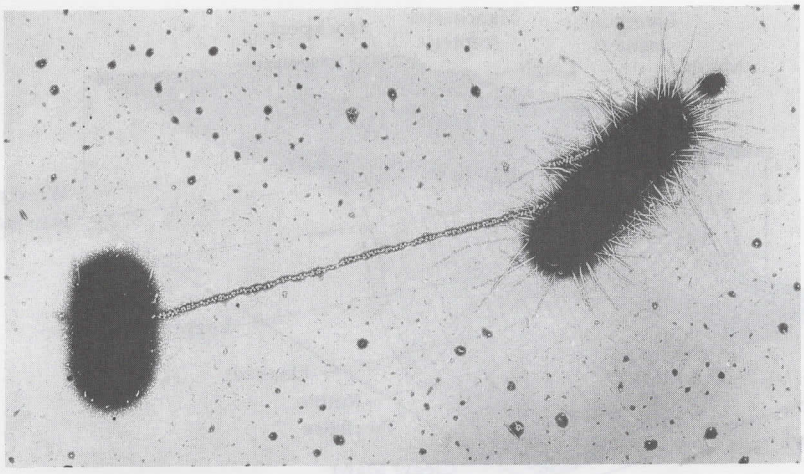


Figure 30. Pilus (the diagonally running structure) connecting a donor *Escherichia coli* cell (upper right) with a recipient *E. coli* cell (lower left). The pilus has been treated with a pilus-specific phage for ease of visualization. This particular type of pilus brings cells into contact for the transfer of a plasmid called the F factor and also for the transmission of certain resistance transfer factors. The numerous other appendages on the donor cell are a different type of pili, not involved in transfer of genes but required for *E. coli* cells to colonize the intestine. (Courtesy of C. C. Brinton Jr. and J. Carnahan.)

sulfonamide, penicillin, streptomycin, tetracycline, chloramphenicol, ampicillin, kanamycin, neomycin, polymyxin, radiomycin, bluensomycin, spectinomycin, viomycin, and gentamycin); resistance to such heavy metal ions as mercury, cobalt, and nickel; and even resistance to ultraviolet radiation (Campbell 1969; Davies and Smith 1978).

Antibiotic resistance presumably evolves in response to the presence of antibiotics in the environment. Antibiotics are chemicals produced by organisms, generally fungi or bacteria, that are inhibitory to the growth of other microorganisms. Although antibiotics have been known since 1928 when Sir Alexander Fleming in England isolated penicillin from the mold *Penicillium*, the industry did not undergo rapid expansion until after the Second World War. Today several thousand types of antibiotics and their chemical derivatives are known, and use of antibiotics has revolutionized the treatment of infectious disease. Indeed, antibiotics have been misused in medicine by their administration to patients with nonspecific ailments or with virus-caused conditions such as the common cold or influenza. Misuse has been particularly brazen in animal husbandry, where antibiotics have been incorporated into livestock feed as "preventatives" against pathogens. As might be expected, the upshot of the continuous presence of antibiotics in the

bacterial environment has been the evolution of antibiotic-resistant strains. The spread of multiple drug resistance in bacteria is evident in Figure 31, which shows the frequency of antibiotic-resistant *Shigella* in Japanese hospital patients. Similar increases in resistance have been noted in bacterial pathogens of domestic animals (Anderson 1969).

It is not known how the resistance genes on resistance transfer factors arise, but there are a number of clues. In the first place, the molecular mechanisms of resistance are typically quite different from those observed in antibiotic-resistant mutants selected directly in plasmid-free bacteria. The laboratory procedure for obtaining antibiotic-resistant mutants is to spread a large number of bacterial cells on solidified agar medium containing the antibiotic, which prevents growth of all the cells except for the tiny minority that have a spontaneously occurring mutation that confers antibiotic resistance. It was once assumed that direct selection provided a suitable experimental model of the evolution of antibiotic resistance in nature. The discovery of resistance transfer factors has shown that the experimental model is much too simple. For example, in the case of streptomycin, the antibiotic exerts its killing effect by interfering with ribosomes and disrupting protein synthesis. Streptomycin-resistant mutants obtained by direct selection have either defective ribosomes that are insensitive to the antibiotic, or altered membranes that prevent entry of the antibiotic into the cell. The spontaneous mutations conferring resistance are themselves harmful, and natural selection acts

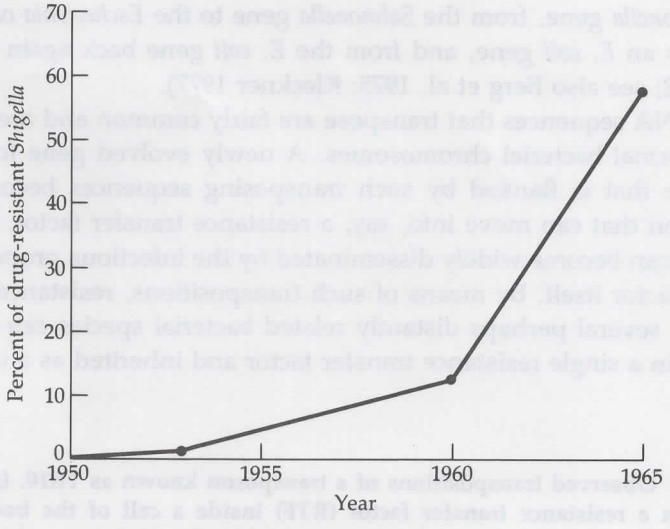


Figure 31. Increase in drug-resistant *Shigella* in Japanese hospital patients during the period 1950 to 1965. The increased drug resistance is due to the spread of resistance transfer factors. (Data from Campbell 1969.)

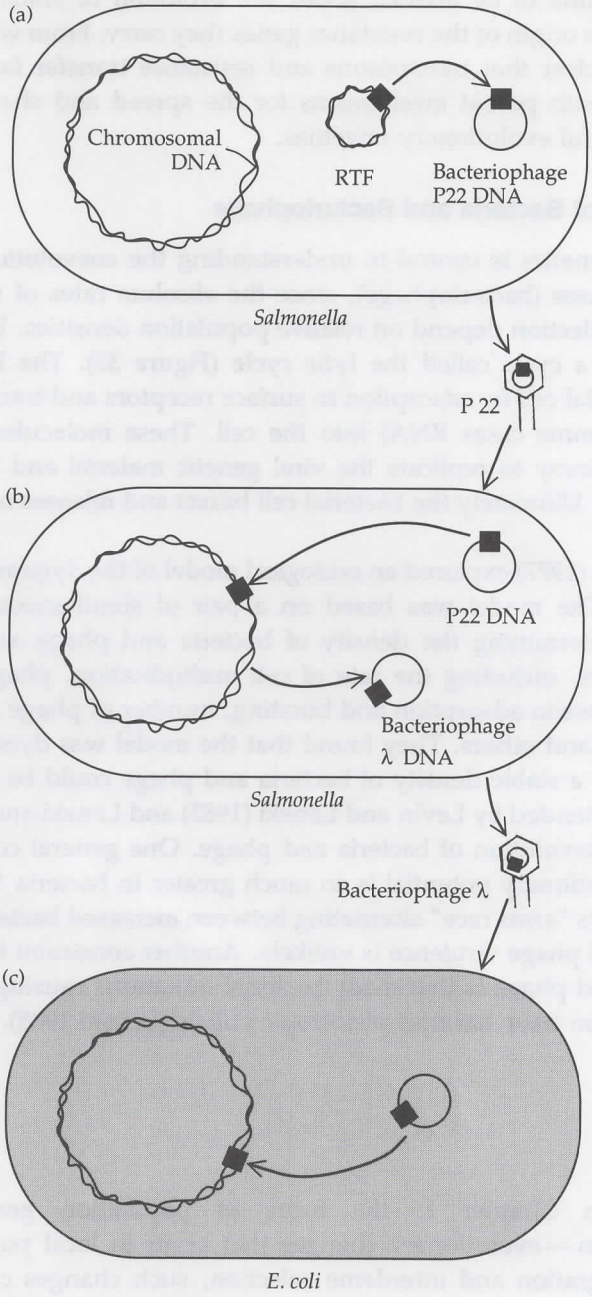
against them in antibiotic-free environments. By contrast, in the case of genes for streptomycin resistance contained in resistance transfer factors, the mechanism of resistance involves neither ribosomes nor membranes, but rather the genes code for enzymes that chemically modify the antibiotic molecule and render it innocuous.

It seems likely that antibiotic-resistance genes evolved in natural populations of bacteria in response to naturally occurring antibiotics. Many microorganisms produce antibiotics, particularly soil-dwelling fungi and bacteria of the genera *Actinomyces* and *Streptomyces* (Benveniste and Davies 1973). For example, in the Solomon Islands, where antibiotics have not been widely used, streptomycin and tetracycline resistance transfer factors have been isolated from both soil and stool specimens (Gardner et al. 1969). It would thus seem that the wide use of antibiotics in medicine and agriculture has simply promoted the spread of resistance transfer factors that probably already existed in nature.

The means by which resistance transfer factors can accumulate genes for resistance to several antibiotics has been clarified somewhat by examination of their molecular structure. The factors contain any of a number of special DNA sequences that can change position from one place in a DNA molecule to another place in the same or a different molecule, often leaving a copy behind as well. Elucidation of the mechanism of transposition is currently a hot topic among molecular biologists. One transposon carrying tetracycline resistance has been followed genetically as it transposed from its original position on a resistance transfer factor to the *Salmonella* phage P22, from P22 to a *Salmonella* gene, from the *Salmonella* gene to the *Escherichia coli* phage λ, from λ to an *E. coli* gene, and from the *E. coli* gene back again to phage λ (Figure 32; see also Berg et al. 1975; Kleckner 1977).

The DNA sequences that transpose are fairly common and are even present in normal bacterial chromosomes. A newly evolved gene for antibiotic resistance that is flanked by such transposing sequences becomes a new transposon that can move into, say, a resistance transfer factor. From there the gene can become widely disseminated by the infectious properties of the transfer factor itself. By means of such transpositions, resistance genes that evolve in several perhaps distantly related bacterial species can be brought together in a single resistance transfer factor and inherited as a unit.

Figure 32. Observed transpositions of a transposon known as Tn*10*. (a) Originally located on a resistance transfer factor (RTF) inside a cell of the bacterial genus *Salmonella*, Tn*10* moved to a P22 phage that had infected the same cell. (A phage, or bacteriophage, is a virus that infects bacterial cells.) (b) When the P22 phage infected a different *Salmonella* cell, Tn*10* moved to the bacterial DNA of this new



host, and from there to a phage λ that had infected the same *Salmonella* cell. The phage λ bearing Tn*10* then infected a cell of the bacterial species *Escherichia coli* (shaded), and the transposon moved onto the *E. coli* chromosome; from there it moved onto the DNA of a different λ bacteriophage. (Data from Kleckner 1977.)

Much remains to be learned about the evolution of resistance transfer factors and the origin of the resistance genes they carry. From what is already known, it is clear that transposons and resistance transfer factors provide prokaryotes with potent mechanisms for the spread and sharing between species of useful evolutionary novelties.

Coevolution of Bacteria and Bacteriophage

Population genetics is central to understanding the coevolution of bacteria and their viruses (bacteriophage), since the absolute rates of mutation and intensity of selection depend on relative population densities. Bacteriophage reproduce in a cycle called the **lytic cycle** (Figure 33). The bacteriophage infect a bacterial cell by adsorption to surface receptors and transferring their DNA (or in some cases RNA) into the cell. These molecules subvert the cellular machinery to replicate the viral genetic material and to synthesize viral proteins. Ultimately the bacterial cell bursts and releases many progeny phage.

Levin et al. (1977) explored an ecological model of the dynamics of bacteria and phage. The model was based on a pair of simultaneous differential equations, determining the density of bacteria and phage as functions of various factors, including the rate of cell multiplication, phage adsorption rate, time between adsorption and bursting, number of phage produced per bacterial cell, and others. They found that the model was dynamically complex, but that a stable density of bacteria and phage could be attained. The model was extended by Levin and Lenski (1983) and Lenski and Levin (1985) to examine coevolution of bacteria and phage. One general conclusion was that the evolutionary potential is so much greater in bacteria than in phage that an endless "arms race" alternating between increased bacterial resistance and increased phage virulence is unlikely. Another constraint in coevolution of bacteria and phage is that many bacterial mutations causing resistance to phage infection have harmful pleiotropic effects (Lenski 1988).

SPECIATION

As noted in Chapter 1, the focus of population genetics is on microevolution — evolutionary changes that occur in local populations. By means of migration and interdeme selection, such changes can be spread throughout an entire species. From the standpoint of population genetics, therefore, the species is perhaps the largest unit of concern. On the other hand, species do split and give rise to new species, and it is the process of

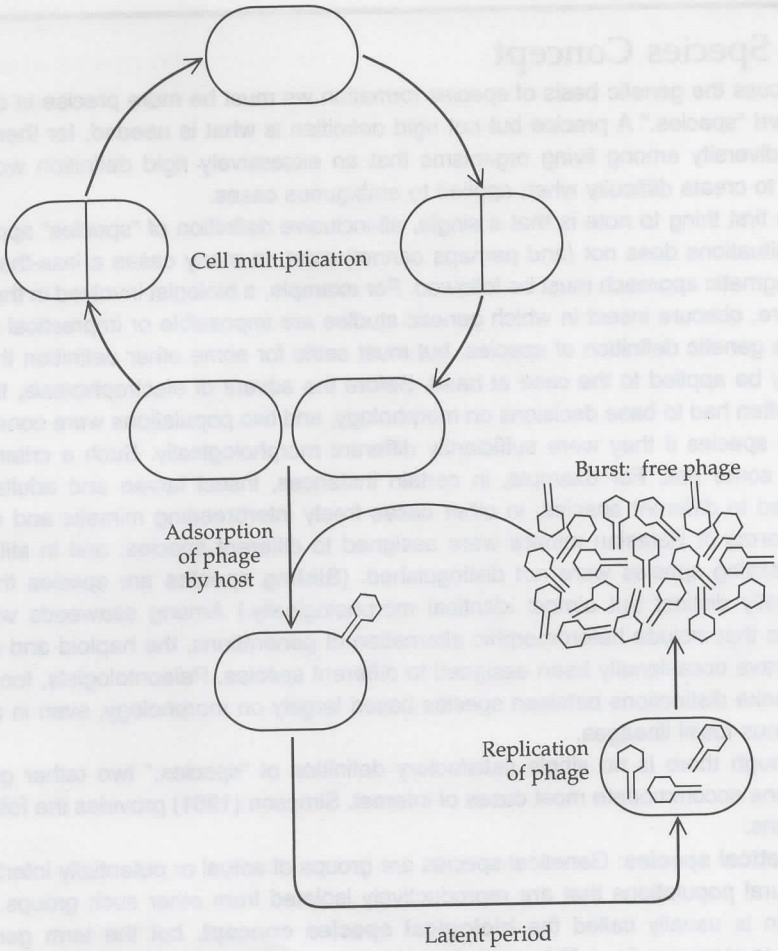


Figure 33. Schematic representation of the dynamics of the lytic cycle of a virulent phage and its bacterial host. (After Levin and Lenski 1983.)

speciation (species formation) that ultimately accounts for the obvious and wide-ranging diversity we see among living organisms. It is generally assumed that the study of population genetics is relevant to understanding the genetic changes that occur during speciation, the argument being that the sorts of genetic change involved in the formation of new species are, at least in some cases, not very different from those that occur in the evolution of a single species. Although the relevance of population genetics to understanding speciation may indeed be profound, the sad truth is that we know almost nothing about the genetics of species formation.

G The Species Concept

To discuss the genetic basis of species formation we must be more precise in defining the word "species." A precise but not rigid definition is what is needed, for there is so much diversity among living organisms that an excessively rigid definition would be bound to create difficulty when applied to ambiguous cases.

The first thing to note is that a single, all-inclusive definition of "species" applicable to all situations does not (and perhaps cannot) exist. In many cases a less-than-ideal but pragmatic approach must be followed. For example, a biologist involved in the study of a rare, obscure insect in which genetic studies are impossible or impractical cannot apply a genetic definition of species, but must settle for some other definition that can actually be applied to the case at hand. Before the advent of electrophoresis, taxonomists often had to base decisions on morphology, and two populations were considered distinct species if they were sufficiently different morphologically. Such a criterion involved some risk. For example, in certain instances, insect larvae and adults were assigned to different species; in other cases freely interbreeding mimetic and nonmimetic forms of Batesian mimics were assigned to different species; and in still other cases sibling species were not distinguished. (**Sibling species** are species that are genetically distinct but almost identical morphologically.) Among seaweeds with life histories that include heteromorphic alternation of generations, the haploid and diploid forms have occasionally been assigned to different species. Paleontologists, too, must often make distinctions between species based largely on morphology, even in almost continuous fossil lineages.

Although there is no single satisfactory definition of "species," two rather general definitions accommodate most cases of interest. Simpson (1961) provides the following definitions.

Genetical species: Genetical species are groups of actual or potentially interbreeding natural populations that are reproductively isolated from other such groups. (This definition is usually called the **biological species concept**, but the term genetical species is more precise.) This concept is the key one in population genetics, and the word "species" will hereafter mean "genetical species" unless otherwise specified. What the above definition states, in essence, is that two populations belong to the same species if they are actually or potentially capable of exchanging genes — that is, if they share a common **gene pool**. Populations that belong to different species do not share a common gene pool by definition and are said to be **reproductively isolated**.

Evolutionary species: An evolutionary species is a lineage (an ancestral-descendant sequence of populations) evolving separately from others and forming a single unit in its evolutionary role. This definition is most useful when genetic studies are impossible or impractical, as is the case in paleontological studies of extinct forms, or when dealing with primarily asexual organisms such as bacteria, or when dealing with preserved museum or herbarium specimens. The word "role" as used in the definition is equivalent to the naturalist's word **niche**, meaning a population's entire way of life and relationship to its environment.

Although the concept of evolutionary species is more general than that of genetical species, the two definitions almost always coincide in sexually reproducing organisms.

The fundamental reason for this correspondence is known to ecologists as the **principle of competitive exclusion**, which states that two distinct species cannot occupy the same niche indefinitely: one species must inevitably change its niche or be driven to extinction by the other (Gause 1934; Hardin 1960). Thus, two coexisting populations that share a common niche will almost always interbreed, and two populations with distinct niches will almost always be reproductively isolated.

Modes of Reproductive Isolation

The establishment of reproductive isolation is obviously a crucial step in the origin of species of sexually reproducing organisms, but unfortunately, almost nothing is known about the details of the process. It is known that existing species are reproductively isolated by any of a variety of mechanisms (usually more than one), and these can be classified broadly according to whether they act prior to mating (**premating isolation**) or afterward (**postmating isolation**). Table 2 provides a typical classification of mechanisms for reproductive isolation. Mechanisms of reproductive isolation found in existing species are themselves products of evolution, however, and they say little about the processes by which they evolved, especially about the early stages.

Experiments have been carried out to determine whether artificial selection can create or amplify behavioral differences between populations and lead to increased premating reproductive isolation. An example in the sibling species *Drosophila pseudoobscura* and *D. persimilis* is provided by Kessler (1966), who mixed males of one species with females of the other and artificially

Table 2. Classification of modes of reproductive isolation.

I. Premating isolating mechanisms (prevention of interspecific crosses)
   A. Seasonal or habitat isolation
   B. Behavioral isolation
   C. Mechanical isolation (incompatibility of reproductive structures)

II. Postmating isolating mechanisms (limitation of success of interspecific crosses)
   A. Gametic mortality
   B. Zygotic mortality
   C. Hybrid inviability
   D. Hybrid sterility (partial or complete)

(Based on Mayr 1970.)

selected a "high-isolation" line by mating animals that did not mate interspecifically to members of their own species and a "low-isolation" line by remating animals that did mate interspecifically to members of their own species. Artificial selection was successful. After 16 generations the low-isolation line produced 22% interspecific matings in an experimental test, and the high-isolation line produced only 2% interspecific matings; both of these values were significantly different from the 10% interspecific matings found in the unselected control populations.

Knight and collaborators (1956) demonstrated that artificial selection for premating reproductive isolation can be effective even within a single species. In experiments with *Drosophila melanogaster*, flies homozygous for either the recessive *ebony body* (*e*) mutation or the recessive *vestigial wing* (*vg*) mutation were mixed, and in each generation the offspring of the $e \times vg$ matings (which are heterozygous for both *e* and *vg* and therefore phenotypically normal) were discarded, thus preserving only the offspring of $e \times e$ and $vg \times vg$ matings. After 20–30 generations, significant premating isolation had occurred, with the selected populations yielding about 39% $e \times vg$ matings compared to 45% in the unselected controls.

Not only is direct selection for premating reproductive isolation successful, but Powell (1978) has shown that it rather frequently arises by chance in populations subject to periodic bottlenecks. Powell studied eight populations of *D. pseudoobscura*, each descended from a single pair of flies. Each population was allowed to expand in size (or "flush") in its own large population cage, and after the flush each population was reestablished from a single pair of flies isolated from the cage. After four founder–flush cycles in each population, Powell found no evidence of postmating reproductive isolation, but three of the eight populations gave evidence of some premating reproductive isolation.

These experiments show that genetic variation affecting reproductive isolation does occur in populations and that at least partial reproductive isolation can evolve in response to artificial selection or, as in the experiments of Powell (1978), by pure chance (see also Thoday and Gibson 1970; and Scharloo 1971). However, we do not know whether artificial selection can push reproductive isolation to completion (that is, to the formation of new genetical species), nor do we know the genetic basis of the response to selection. The issues are amenable to experimental analysis and deserve much more attention than they have received.

Geographical Aspects of Speciation

The key element in speciation among sexual organisms is the evolution of reproductive isolation, or mechanisms that restrict or prevent gene flow

(migration) between populations. Geographical barriers, such as rivers or mountains, often suffice to prevent gene flow between populations, and such geographical effects are the basis of **allopatric speciation**, a mode of speciation mediated by physical separation of the populations undergoing speciation (Mayr 1963, 1970). Alternative modes of speciation with no pronounced geographical separation are also known. In **parapatric speciation** (White 1968, 1974), populations that are contiguous undergo the evolution of reproductive isolation; in **sympatric speciation** (Maynard Smith 1966), a new species arises entirely within the geographical range of its parental form (for recent reviews see Bush 1975; Carson 1975, 1978; White 1978; Templeton 1980; Carson and Templeton 1984). Allopatric, parapatric, and sympatric speciation are the principal modes of species formation; although more detailed classifications are useful for some purposes (White 1978), we will focus on these three broad categories.

Allopatric speciation

There are actually two types of allopatric speciation, differing somewhat in their genetic concomitants. The two types are distinguished by whether or not a marked founder effect or population bottleneck is involved.

Type Ia allopatric speciation (in the terminology of Bush 1975) can occur when a geographical barrier (a river, desert, mountain range, glacial tongue, or any number of things) splits a species with a wide geographical range into two or more large subpopulations between which gene flow is severely restricted (Figure 34a). In such a situation, each geographically isolated subpopulation is free to evolve in its own way, adapting to its own local conditions. Although the initial large sizes of the subpopulations insure that each will carry essentially all the genetic variation present in the original population, the subpopulations will diverge genetically, each evolving toward an adaptive peak appropriate to its local conditions. While the subpopulations are geographically isolated, there is no direct selection for reproductive isolation. Reproductive isolation could evolve as a correlated response to natural selection for other characteristics, or genetic differentiation could become so pronounced that hybrid genotypes would be selectively inferior. In the latter case, should the subpopulations again come into contact, there would be selection in favor of premating reproductive barriers (this is called the **Wallace effect**), because the organisms that hybridize produce inferior offspring (Sawyer and Hartl 1981). Type Ia allopatric speciation is apparently a slow process, and the sorts of genetic changes that occur in populations undergoing speciation are the same sorts of microevolutionary changes that occur in any large evolving population.

Type Ib allopatric speciation (Figure 34b) is geographically similar to Type Ia, but Type Ib includes a pronounced founder effect due to a small number

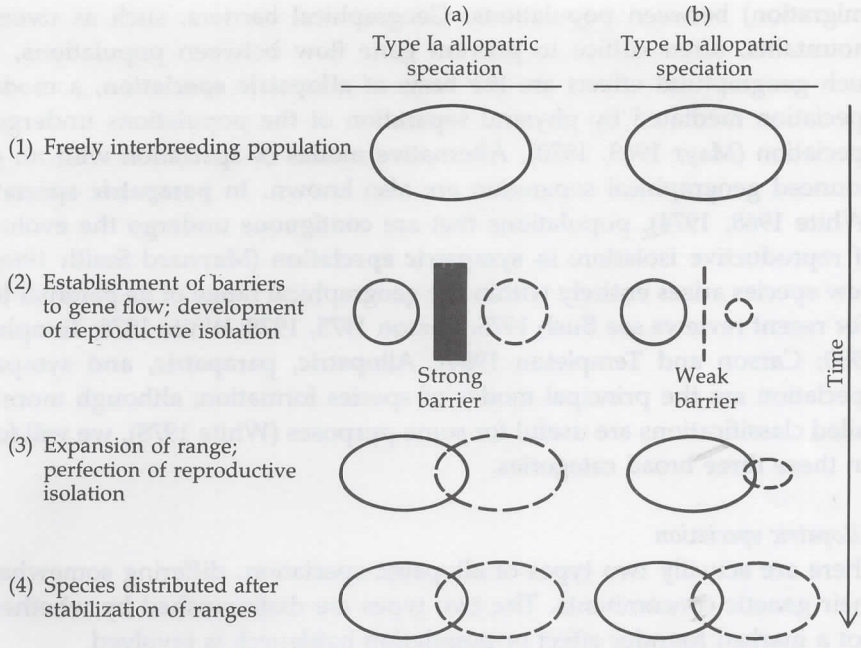


Figure 34. (a) Type Ia allopatric speciation, in which a strong geographical barrier to gene flow splits a population (1) into two large subpopulations (2) that evolve independently and develop incipient reproductive isolation. Reproductive isolation is perfected should the barrier than disappear or become surmounted (3), leading to overlapping species ranges (4). (b) Type Ib allopatric speciation, in which a small subpopulation (with consequent large founder effects) splits off from a large population (2), and, by random genetic drift, develops incipient reproductive isolation; the populations thereafter follow patterns (3) and (4) of Type Ia allopatric speciation. (After Bush 1975.)

of propagules fortuitously founding a new colony geographically isolated from the parental population. The initial colony will be small, perhaps consisting of a single fertilized female, and the resulting population will therefore experience extensive inbreeding and random genetic drift (Carson 1968, 1971; Templeton 1980). Indeed, because of the change in allele frequencies, the population may evolve to a new adaptive peak even in the absence of an environmental change. As with Type Ia allopatric speciation, initial reproductive barriers are assumed to arise by chance and to be perfected first in the area of overlap should the populations regain contact.

Allopatric speciation is generally agreed to be the principal mode of speciation in most groups of plants and animals. Type Ia appears to be characteristic of most outbreeding animals of high vagility that typically have low reproductive rates and exist at population densities near the carrying capacity

of the environment. (**Vagility** refers to the ability of organisms to disperse.) Included here are large mammals such as felids (cats) and canids (dogs), most birds, certain reptiles and amphibians, and most large fish. Social organization into family groups, harems, or herds promotes Type Ib speciation, especially in organisms of low vagility such as primates. High vagility in organisms that exist in patchy environments also seems to promote speciation according to the Type Ib mode.

Allopatric speciation has been particularly well-studied in the Hawaiian Drosophilidae, where 400 endemic drosophilid species have been described (the total number of endemic species is estimated at 700 or more). Although the Hawaiian Islands that exist today are geologically quite young (Kauai, the oldest island, is about 5.6 million years old, and Hawaii, the youngest, is about 1 million years old), the flies themselves have been evolving and undergoing adaptive radiation for about 42 million years (Beverly and Wilson 1985). This apparent paradox is explained by the geological origin and history of the archipelago. Deep beneath the present island of Hawaii is a hot spot in the Earth's mantle that creates volcanic islands that swell above the surface, persist for some time, and then erode or submerge. As the sea floor has moved slowly across the hot spot in a northwesterly direction, an extensive archipelago formed, which today consists of a series of undersea mountains and plateaus, coral reefs, and atolls arching across more than 6000 km, from south of the Komandorskiye (Commander) Islands in the Bering Sea to the Koko seamount (created by the Hawaiian hotspot about 46 million years ago), through Midway (created about 30 million years ago), to present-day Hawaii.

Based on molecular comparisons of the larval hemolymph of Hawaiian and other drosophilids, Beverly and Wilson (1985) inferred that drosophilids originally colonized the long archipelago some 42 million years ago. As the chain of islands rose and fell and crept slowly across the Pacific, the drosophilids evolved and diversified, migrating from island to island (Figure 35). While today's Hawaiian drosophilids consist of hundreds of endemic species, and the islands themselves have existed for only a few million years, the total evolutionary divergence between the Hawaiian drosophilids and other drosophilids in the archipelago indicates a long period of previous evolution as the flies island-hopped along the slowly moving archipelago.

The endemic Hawaiian drosophilids are extraordinarily diverse in morphology and sexual behavior. Moreover, since the Hawaiian Islands lack native fleshy fruits like those that provide food and shelter for most continental species, the Hawaiian species have evolved highly specialized and sometimes bizarre feeding habits, such as feeding on slime fluxes from trees, decaying bark, or even on spider eggs. Speciation in the Hawaiian drosophilids has been studied extensively by Carson and colleagues (Carson 1971,

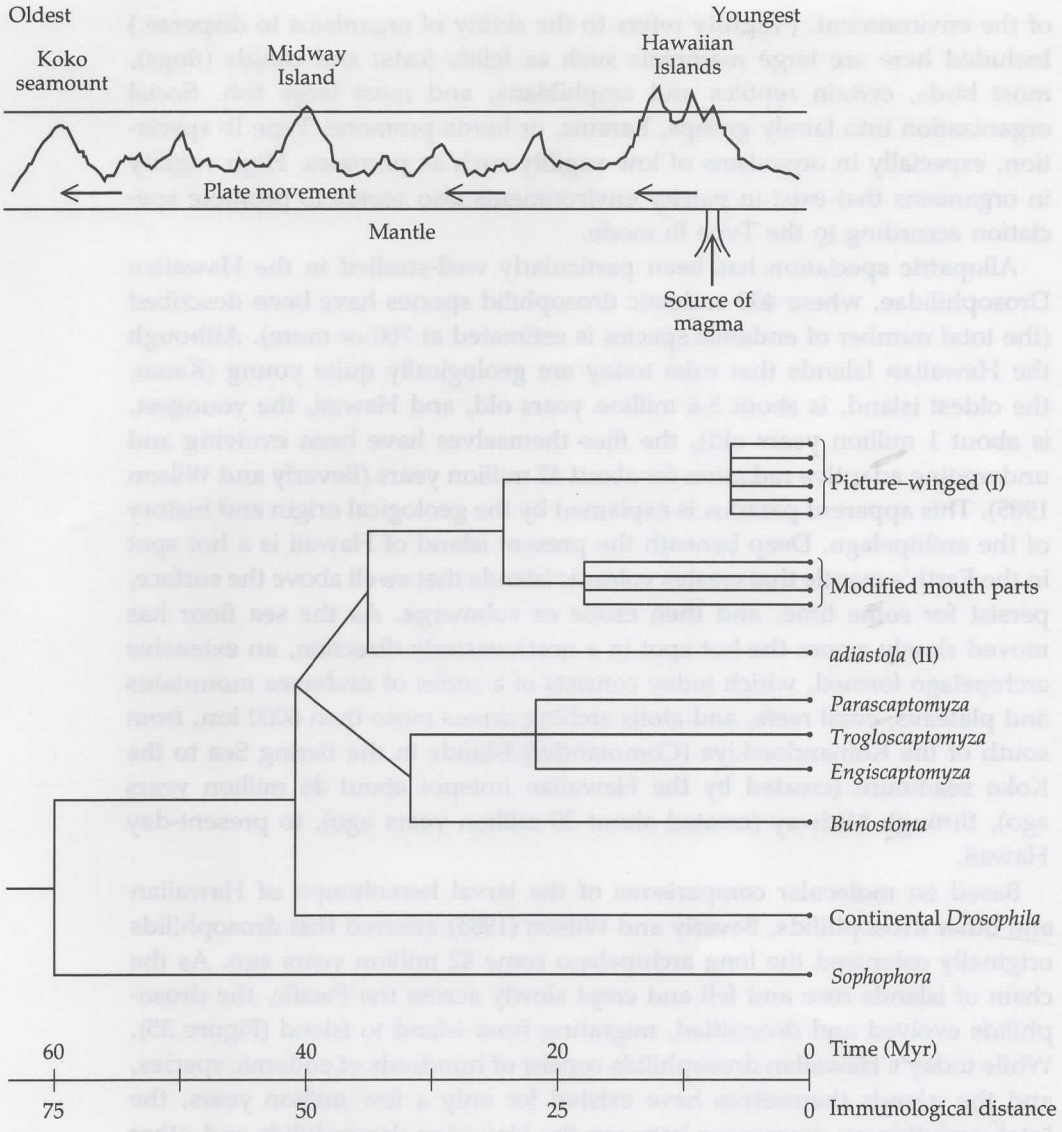


Figure 35. Molecular tree for Hawaiian drosophilid flies in relation to time and geography. The tree is based on an immunological distance scale using the larval hemolymph proteins of the Hawaiian species. Continental *Drosophila* refers to the Hirtodrosophila, *immigrans*, *pinicola*, *repleta*, *virilis*, and *robusta* groups; *Sophophora* includes members of the *melanogaster*, *subobscura*, *willistoni*, and *saltans* groups. The time scale above is empirically related to the immunological distance scale as described in Chapter 7. Above the tree is a cross-section of the Hawaiian archipelago, showing how the movement of the crustal plate over the stationary magma source resulted in a chain of islands and submerged mountains.

1975, 1978; Carson and Kaneshiro 1976; Carson and Bryant 1979), who have argued that the speciation may be promoted by repeated cycles of population expansion (population flush) under favorable environmental conditions when selection is relaxed (Carson 1968). The large population and relaxed selection at the time of the flush also allows new recombinant genotypes to be formed. Such conditions may permit more genotypes to survive and reproduce than would at other times and, since emigration pressure is greatest at the peak of the flush when the population is most dense, there is enhanced likelihood that founders of new colonies will be genotypically quite distinct from the parental population. Speciation by this sort of **founder–flush model** is a special case of Type Ib speciation, and the finding of Powell (1978) that founder–flush cycles often lead to incipient reproductive isolation lends credence to the hypothesis.

Although allopatric speciation is usually considered to be mediated by genetic changes during the period of isolation, extrachromosomal factors can also form an important postmating barrier. Several forms of maternally transmitted factors have been implicated in differential success of the reciprocal F_1 hybrids of several species (Hoffman et al. 1986). In the case of *Drosophila simulans*, there is variation among populations in California in cytoplasmic types that determines the success of hybridization. This unidirectional hybrid incompatibility can be rendered benign by administering tetracycline, and appears to be mediated by a cytoplasmic microorganism (Hoffman and Turelli 1988).

Speciation by the Type Ib mechanism can be rapid, in contrast to Type Ia, which tends to be slow. For example, among the cichlid fish of the Great Lakes of East Africa, five species are endemic to Lake Nabugabo, though the lake is only 3500 years old, and more than 150 species are endemic to 750,000-year-old Lake Victoria (Fryer and Iles 1972; discussed in Bush 1975).

Parapatric speciation

Contiguous populations sometimes undergo speciation, especially those of organisms with low vagility that exist in numerous small subpopulations; such parapatric speciation has been studied in the flightless morabine grasshoppers of Australia (White 1969), snails of the genus *Partula* (Clarke and Murray 1969), the burrowing mole rats of Israel (Nevo and Cleve 1978), and plants that have limited pollen dispersal, such as those cited earlier in connection with heavy metal tolerance.

In the case of parapatric speciation, low vagility combined with small subpopulations serves to accentuate the effects of random genetic drift and promotes genetic differentiation. Postmating reproductive isolation due to hybrid inviability or infertility can arise rapidly, especially if chromosomal

rearrangements, which may become fixed in small populations due to chance, are involved. In the approximately 200 species of flightless grasshoppers of the subfamily Morabinae, for example, 39 chromosomal fusions (due to translocations) and 22 dissociations (due to chromosome breakage) are known to have occurred, and most species are chromosomally unique. Zones of overlap between morabine species are invariably narrow; hybridization does occur in these zones, but gene flow between the species is essentially eliminated by the low fertility of the chromosomally abnormal hybrids (White 1978).

Sympatric speciation

In certain cases, new species can evolve well within the dispersal range of the parental species. This mode of speciation is sympatric speciation, and it is apparently limited to certain special types of organisms such as insect parasites, especially those that are host-specific and in which courtship and mating occur on or in the host. In a population of this sort, any individuals with an altered host specificity will court each other and mate while associated with their special host, and they will in this manner be reproductively isolated from the parental population, because a premating reproductive barrier is brought about by the distinct host preference. Under these circumstances there is no strong selection for premating reproductive isolation, and in at least one case, that of the European cherry fruit flies of the genus *Rhagoletis* (Boller and Bush 1973), species can be hybridized in the laboratory and yield fully fertile offspring. This does not violate the biological species concept because the two species do not encounter one another in nature.

The number of genes responsible for host specificity may be few — for example, those concerned with chemoreception. A single gene controls recognition of host plants in the gall-forming fruit fly *Procecidochares* (Huettel and Bush 1972). Moreover, survival on a new host need not be very good at first, since subsequent evolution of a conventional sort will progressively adapt the new species to life on its new host.

With sympatric speciation, therefore, a new species can arise almost instantaneously and may differ from the parental species in only a few genes. Another virtually instantaneous mode of speciation often occurs sympatrically or parapatrically, which is of fundamental importance in the evolution of plants, especially higher plants. This mode occurs by means of the formation of fertile autopolyploids or allopolyploids, and its importance in plant evolution is well established (Stebbins 1950). Polyploidy results in instantaneous postmating reproductive isolation because the interspecific hybrids are chromosomally unbalanced and highly infertile. (There may also be subsequent selection for premating reproductive isolation.) Although autopolyploids would seem to be genetically much the same as their diploid ancestors,

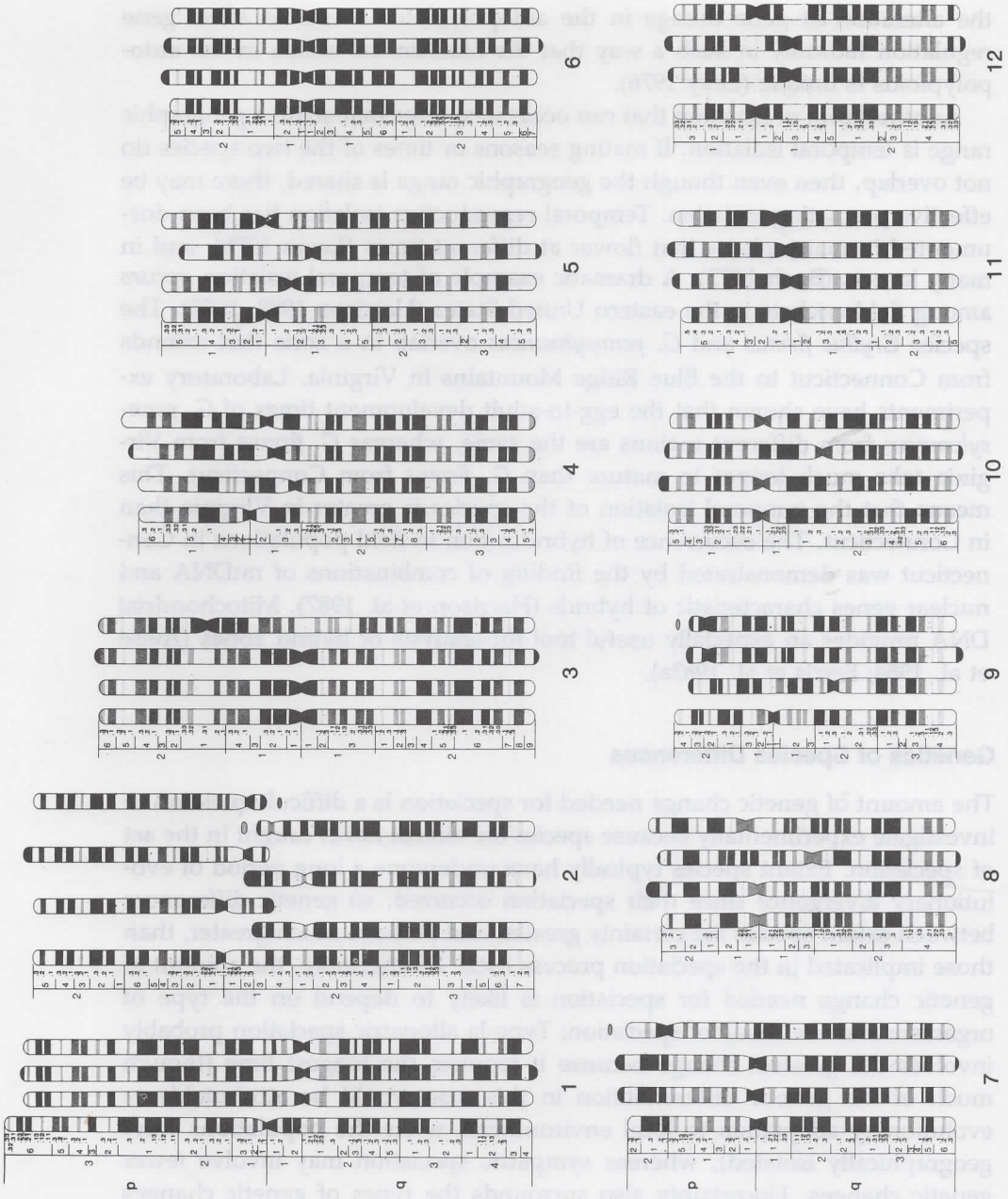
the alteration of gene dosage in the autopolyploids seems to affect gene regulation radically in such a way that the biochemical milieu in the autopolyploids is unique (Levy 1976).

Another form of isolation that can occur in species that share a geographic range is temporal isolation. If mating seasons or times of the two species do not overlap, then even though the geographic range is shared, there may be effective premating isolation. Temporal reproductive isolation has been documented in many plants that flower at different times (Levin 1978), and in many insects (Bush 1975). A dramatic example of temporal isolation occurs among field crickets in the eastern United States (Harrison 1983, 1985). The species *Gryllus firmus* and *G. pennsylvanicus* overlap in a zone that extends from Connecticut to the Blue Ridge Mountains in Virginia. Laboratory experiments have shown that the egg-to-adult development times of *G. pennsylvanicus* from different regions are the same, whereas *G. firmus* from Virginia take much longer to mature than *G. firmus* from Connecticut. This means that the temporal isolation of the species is greater in Virginia than in Connecticut. The occurrence of hybridization in field populations in Connecticut was demonstrated by the finding of combinations of mtDNA and nuclear genes characteristic of hybrids (Harrison et al. 1987). Mitochondrial DNA provides an especially useful tool for analysis of hybrid zones (Avise et al. 1984; Ferris et al. 1983a).

Genetics of Species Differences

The amount of genetic change needed for speciation is a difficult question to investigate experimentally because species are almost never caught in the act of speciation. Extant species typically have undergone a long period of evolutionary divergence since their speciation occurred, so genetic differences between extant species are certainly greater, and perhaps much greater, than those implicated in the speciation process itself. Furthermore, the amount of genetic change needed for speciation is likely to depend on the type of organism and the mode of speciation; Type Ia allopatric speciation probably involves the greatest change because it requires the longest time (though much of the genetic differentiation in this case would be attributable to evolutionary adaptation to local environments while the populations were geographically isolated), whereas sympatric speciation may involve fewer genetic changes. Uncertainty also surrounds the types of genetic changes that occur during speciation — specifically, whether they primarily involve genetic changes at the molecular level or visible changes in chromosome structure.

The **karyotype**, or organization of genetic material into chromosomes, is


1
2
3
4
5
6
7
8
9
10
11
12
p
q

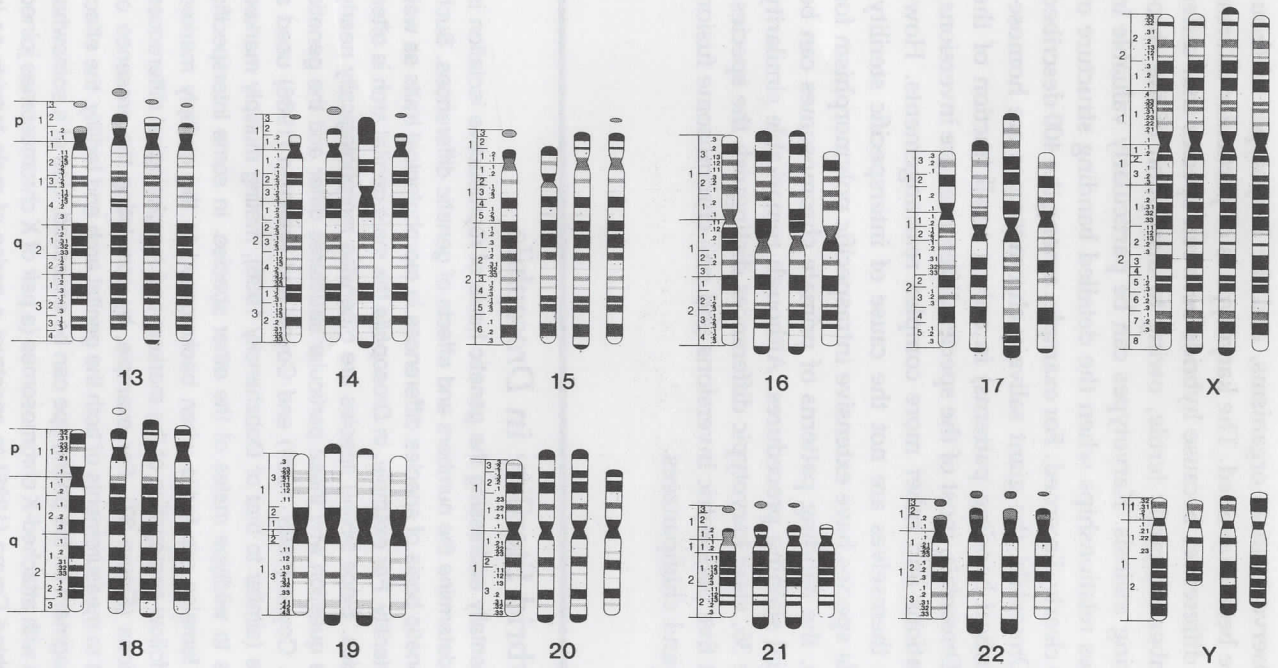


Figure 36. Schematic drawings of the G-banded karyotypes of human, chimpanzee, gorilla, and orangutan (left to right, respectively), arranged so that the homology of banding patterns is evident. (From Yunis and Prakash 1982, © 1982 by the AAAS.)

relatively easy to observe in most organisms, and the karyotypes of thousands of species have been examined. The karyotype is of particular interest in regard to species differences because hybrids between species that differ in chromosome number are rarely fertile, owing to disruption of the chromosome pairing during meiosis. Karyotypes can be particularly valuable in understanding species relationships when the detailed banding structure of chromosomes can be clearly discerned. For example, among the 400 described endemic Hawaiian *Drosophila*, the giant salivary chromosomes are homosequential (having identical banding patterns) in only a small fraction of the species. As in other *Drosophila*, most of the species differences are inversions, rather than translocations and other more complex rearrangements. However, the inversions themselves are not the cause of interspecific sterility, since many *Drosophila* species have extensive intraspecific polymorphism for inversions. Likewise, the banding patterns of primate chromosomes can be compared with special staining procedures. Although remarkable similarity is apparent in Figure 36, clear karyotypic differences distinguish the species. For example, at least five pericentric inversions and one chromosome fusion distinguish humans and chimpanzees.

H Genetics of Hybrid Sterility in *Drosophila*

One means of experimentally addressing the genetic basis of reproductive isolation is to hybridize species to determine the numbers and effects of genetic differences. Such crosses address the genetic basis of species differences in morphological traits as well as interspecific hybrid sterility. For example, in *Drosophila* the male genital arch is often different in sibling species. Since sibling species are otherwise morphologically nearly identical, this raises the question why these particular structures differ and the genetic basis of the difference. Coyne (1983, 1984) and Coyne and Kreitman (1986) used a simple crossing scheme (similar to that of Dobzhansky 1936), mating multiply marked females of one species to wildtype males of the other species. In some interspecific crosses the F_1 hybrid females are fertile when backcrossed to the multiply marked males, enabling one to follow segregation of the markers and morphological differences in subsequent generations (Figure 37). For example, by correlating the presence of chromosomal segments to measurements of both the genital arch and fertility, the effect of each chromosomal segment on the phenotype can be assessed. Use of a somewhat more elaborate scheme with attached-X chromosomes (a pair of X chromosomes joined at the centromere) enabled Coyne (1985) to construct a series of male hybrids as in Figure 37 with only the Y chromosome from one or the other original species. In tests with *Drosophila simulans*, *D. mauritiana*, and *D. sechellia*, he showed that male sterility was due to incompatibilities between the X and the Y chromosome, thereby establishing a genetic basis of Haldane's rule, which states that if only one sex of the F_1 hybrids is

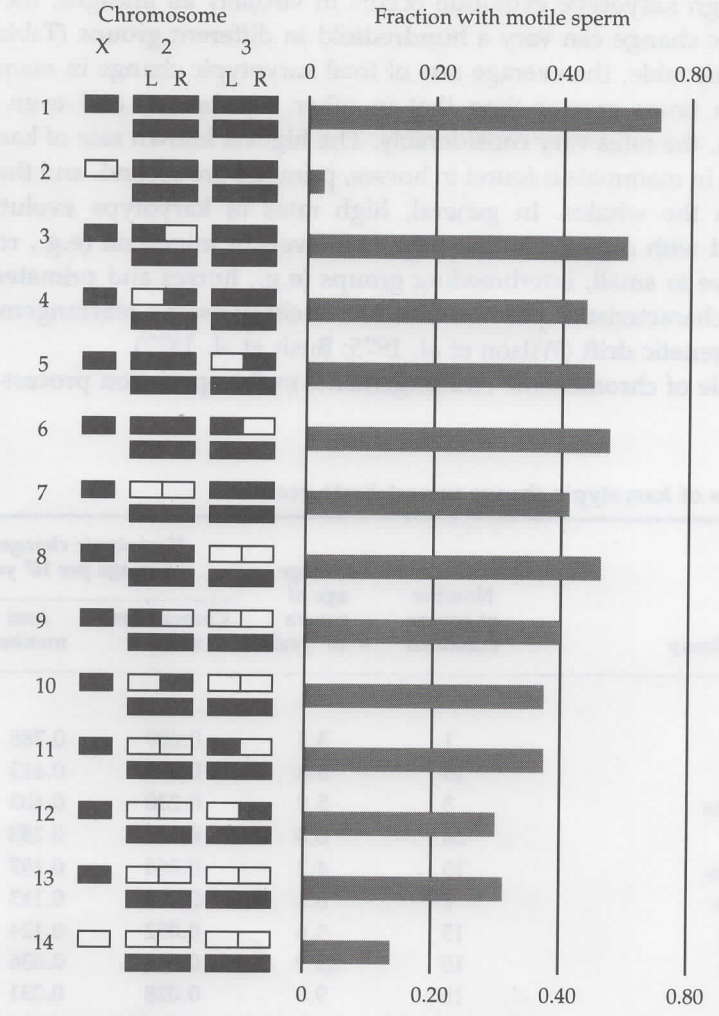


Figure 37. Karyotypes of various backcrossed interspecific hybrids of *Drosophila simulans* and *D. sechellia* and their resulting sperm motility. Segments of the *D. simulans* chromosomes are shown in black, and those of *D. sechellia* are shown in white. All males have the *D. simulans* Y chromosome, and the magnitude of the X–Y incompatibility can be seen in the low sperm motility of the second genotype from the top. (From Coyne and Kreitman 1986.)

sterile, it is the heterogametic sex. Just as the morphological studies suggest relatively few genetic differences between the species, analysis of *Adh* sequences of the same species reveals that there is as much variation within species as between species (Coyne and Kreitman 1986). This again emphasizes the pitfalls of inferring phylogenetic relationships using single sequences from each taxon.

Although karyotype evolution occurs in virtually all lineages, the rate of karyotypic change can vary a hundredfold in different groups (Table 3). As seen in this table, the average rate of total karyotypic change in mammals is about ten times greater than that in other vertebrates, and even among mammals, the rates vary considerably. The highest known rate of karyotypic evolution in mammals is found in horses, primates are second, and the lowest rate is in the whales. In general, high rates of karyotype evolution are associated with organisms that have low levels of migration (e.g., rodents), or that live in small, interbreeding groups (e.g., horses and primates). Both of these characteristics promote fixation of chromosome rearrangements by random genetic drift (Wilson et al. 1975; Bush et al. 1977).

The role of chromosome rearrangements in the speciation process itself is

Table 3. Rates of karyotypic change in vertebrate genera.

Group	Number of genera examined	Average age of genera × 10^6 years	Karyotypic changes per lineage per 10^6 years		
			Chromosome number	Arm number	Total
Mammals					
Horses	1	3.5	0.609	0.786	1.395
Primates	13	3.8	0.333	0.413	0.746
Lagomorphs	3	5.0	0.230	0.403	0.633
Rodents	50	6.0	0.178	0.253	0.431
Artiodactyls	15	4.2	0.364	0.197	0.561
Insectivores	7	8.1	0.074	0.113	0.187
Marsupials	15	5.6	0.052	0.124	0.176
Carnivores	10	12.9	0.042	0.036	0.078
Bats	15	9.0	0.028	0.031	0.059
Whales	2	6.5	0.000	0.025	0.025
Average for mammals		6.5	0.129	0.166	0.295
Other vertebrates					
Lizards	16	20.1	0.027	0.031	0.058
Snakes	14	12.1	0.007	0.041	0.048
Turtles and crocodiles	14	45.2	0.0006	0.0016	0.002
Frogs	15	26.4	0.011	0.012	0.023
Salamanders	11	23.4	0.006	0.008	0.014
Teleosts	12	5.7	0.003	0.026	0.029
Average for nonmammals		22.1	0.009	0.020	0.029

(From Bush et al. 1977.)

still unclear, however. In groups that undergo speciation predominantly according to the Type Ib allopatric model or according to the parapatric model, chromosomal differentiation owing to founder effects might be common or even crucial to the origin of postmating reproductive isolation. In such groups as frogs and whales, which are thought to undergo speciation primarily according to the Type Ia allopatric model, karyotypic changes that arise during speciation would be expected to be less frequent.

A good example of karyotype evolution occurs among the salmonid fishes. Fish in the family Salmonidae (trout, char, and salmon) have twice the number of chromosome arms and twice the DNA content per cell as their close relatives, the smelts and herrings. The Salmonidae appear to have descended from a tetraploid ancestor (reviewed in Wright et al. 1983). In the brook trout, the chromosome number is $2n = 84$; although females show the expected 42 bivalents (paired chromosomes in meiosis), males show variable numbers of tetravalents, reminiscent of tetraploids (Lee and Wright 1981). The tetravalents help explain the phenomenon of "pseudolinkage," in which markers on different chromosomes appear linked (Johnson et al. 1987). As the pairs of chromosomes that were once homologous accumulate differences, the tetravalent pairing in males becomes less frequent. This process, known as *diploidization*, provides an excellent opportunity to study one genetic aspect of speciation as it occurs.

The Bizarre Genetics of *Poeciliopsis*

I

Poeciliopsis is a small fish that inhabits freshwater streams in Mexico; there are several races of *Poeciliopsis* that exhibit unusual modes of reproduction. Some races are all-female triploid clones that reproduce by gynogenesis. This in itself is not so unusual, but the triploid eggs require sperm from other races for the initiation of development (Schultz 1969; Vrijenhoek 1972, 1978). Other races reproduce by hybridogenesis, in which diploid females (with one haploid complement of *Poeciliopsis monacha* and one haploid complement of *P. lucida*, designated ML) produce haploid eggs with only the M genome, and these are fertilized by sperm from the sexual species *P. lucida* (LL), recreating the ML hybrid that will in turn transmit only M gametes. (The hybridogens are also called hemiclones, since the M genome is passed clonally but the L genome is not.)

Some streams contain the normal diploid sexual species *P. lucida* and *P. monacha* (the source of the M genome) coexisting with several triploid gynogen clones and several diploid hybridogen clones. Statistical analysis of fish populations and recorded environmental data suggest that the distribution of the various clones of fish is determined by ecological factors (Schenk and Vrijenhoek 1986). The occurrence of the clonally reproducing forms allows the possibility of "freezing" genotypes ideally suited to particular habitats. However, the requirement for hybrid mating of the hybridogens challenges usual notions of species identity and the importance of premating barriers.

In some cases even a single locus can have a major impact on species isolation. When *D. melanogaster* females are inseminated by *D. simulans* males, normally only sterile female offspring are produced, but if the male *D. simulans* has the mutation *Lethal hybrid rescue* (*Lhr*), sterile male offspring are produced as well (Watanabe 1979). A similar phenomenon has been found in hybrids of *D. mojavensis* and *D. arizonensis*, whose sperm motility seems to depend on a single autosomal Mendelian factor (Pantazidis and Zouros 1988). Once a barrier to gene flow exists, genetic differences will continue to accumulate, but the complexity of the genetic basis for the initial isolation need not be continuously buttressed.

Species can differ at all levels of genetic organization. DNA sequence differences can result in synonymous or nonsynonymous substitutions in structural genes, they may affect gene expression by changes in polymerase binding sites or exon–intron splice junctions, they may result in changes in copy number of repeated genes, and so on. Despite the abundant sequence data currently available, only a tiny portion is useful for examining genetic differences involved in speciation. Likewise the many population genetic studies of intra- and interspecific genetic variation using protein electrophoresis have been useful in estimating the magnitude of species differences, but these approaches are fundamentally descriptive and do not directly address the genetic mechanisms of speciation. The genetic basis of speciation will therefore remain one of the most challenging problems in evolutionary genetics.

PROBLEMS

1 In the evolution of industrial melanism, which would occur more rapidly, selection for an initially rare dominant melanic allele or selection for a recessive? After the melanic allele becomes common, which case could be reversed more rapidly in an unpolluted environment?

2 What kind of genotype would be favored in an unpredictable environment fluctuating between marginal and optimal conditions, one with a high intrinsic rate of increase or one with a high carrying capacity, and why?

3 Why should the repeated observation of similar clines in similar environmental conditions on different continents suggest the operation of natural selection?

4 Why is sympatric speciation much less likely to be common than allopatric speciation?

5 Why is it easier to imagine altruistic traits evolving by group selection when the members of the groups are related?

6 Why is it generally discouraged to use routine antibiotic treatment as a prophylactic against infection?

7 What two features of bacterial resistance transfer factors make them formidable in the fight against microbial disease?

8 In a haploid population, what is the difference if an allele is lethal in half the microenvironments in every generation or in all the microenvironments every other generation?

9 Distinguish between Batesian mimicry and Müllerian mimicry. What kind of frequency-dependent selection is implicit in each process?

10 Which is more effective in changing allele frequency, a selection coefficient of 1% every generation or one of 10% every tenth generation? Demonstrate the situation using a haploid model.

11 In a species (such as wasps) in which males are haploid and result from the development of unfertilized eggs, what is the probability that a locus in a female will have both alleles identical by descent with those in her sister? Assume random mating and that each female mates with only one male.

12 Other things being equal, which will grow faster, a population in which reproduction occurs early in the life cycle or one in which reproduction is delayed?

13 In a steady state metabolic pathway, the control coefficient of the ith step is defined as $C_i^J = (\partial J/J)/(\partial E_i/E_i)$, where J is the flux and E_i the activity of the ith enzyme. Under a wide variety of conditions, it can be shown that $\Sigma C_i^J = 1$. How does this summation theorem demonstrate a limitation to the concept of rate-limiting steps in metabolic pathways?

14 Metabolic control theory suggests a hyperbolic relation between fitness and enzyme activity in cases where fitness is proportional to flux. If fitness is given by the hyperbolic relation $w = a/(1 + a)$, what increase in fitness results from a 10% increase in a when $a = 10$? From a 10% increase in a when $a = 100$?

15 The expression $c/b < r$ specifies the maximum ratio of cost to benefit for the evolution of altruism. When there is inbreeding, r is defined as $2F_{XY}/(1 + F_X)$, where individual X is the altruist, Y is the recipient of the altruism, F_{XY} is the inbreeding coefficient of a hypothetical offspring of X and Y, and F_X is the inbreeding coefficient of X. What does the ratio $2F_{XY}/(1 + F_X)$ mean in words?

16 If an allele a in a haploid population is lethal in a fraction s of environments but neutral otherwise, and the $A \rightarrow a$ mutation rate is μ per generation, what is the equilibrium frequency q of a?

17 For a locus with two alleles in a random mating population, what is the

ultimate fate if the alternate alleles are recessive lethals in alternate generations?

18 If a population contains a normal Y chromosome at frequency y_1 and a meiotic drive Y chromosome at frequency y_2 that results in a fraction $k > 1/2$ of male offspring, show that the frequency of the meiotic drive Y in generation t is given by $y_2(t)/y_1(t) = [y_2(0)/y_1(0)](2k)^t$. What is the ultimate fate of the meiotic drive Y? What fitness reduction in meiotic drive males will exactly offset the meiotic drive and result in no frequency change?

19 When population growth is governed by the logistic equation $dN(t)/dt = rN[1 - (N/K)]$, what is the fitness of a genotype with growth parameters r^* and K^* when the population as a whole is growing with parameters r and K?

20 The Wallace effect refers to selection for premating reproductive isolation among species that have some postmating reproductive isolation. If two species A and B produce inviable or sterile hybrids, then an allele in A with frequency p that causes a premating aversion to species B will increase in frequency approximately according to $\Delta p = pq(1 - \gamma)r$, where r is the amount of mating aversion in heterozygotes (assuming additive effects with $r < 0.4$) and γ is the proportion of species A in the zone of overlap (Sawyer and Hartl 1981). Why does the change in frequency depend on γ? If $\gamma = 1/2$, how does selection for premating isolation compare with viability selection for an additive allele?

Suggestions for Further Reading

The following books are suggested for further reading:

Bulmer, M. G. 1985. *The Mathematical Theory of Quantitative Genetics*. Clarendon Press, Oxford. Along with Falconer and Mather and Jinks, this book presents the basic principles of quantitative genetics. Emphasis is on animal breeding, with perhaps more concern for evolutionary questions than the other texts.

Cavalli-Sforza, L. L. and W. F. Bodmer. 1971. *The Genetics of Human Populations*. W. H. Freeman and Co., San Francisco. Despite the copyright date, this remains one of the best books on the population genetics of humans, and is full of excellent examples.

Charlesworth, B. 1980. *Evolution in Age-Structured Populations*. Cambridge University Press, Cambridge. With an entire book on this subject, it is not surprising that the treatment is advanced and very thorough.

Crow, J. F. and M. Kimura. 1970. *An Introduction to Population Genetics Theory*. Harper & Row, New York. This is the book from which the previous generation learned population genetics theory, and it remains the best source for many aspects of population genetics theory. It was reprinted in 1978 by Burgess.

Ewens, W. J. 1979. *Mathematical Population Genetics*. Springer-Verlag, Berlin. This is a rather advanced text, but covers much of the more recent material. It gives a dense and rigorous treatment of the subject.

Falconer, D. S. 1981. *Quantitative Genetics*, 2nd Ed. Longman, New York. This is the classic text on quantitative genetics, and provides an excellent introduction to the subject.

Feldman, M. W. and F. B. Christiansen. 1986. *Population Genetics*. Blackwell Scientific Publications, Palo Alto, California. An elementary text with emphasis on human examples.

Futuyma, D. J. and M. Slatkin. 1983. *Coevolution*. Sinauer Associates, Sunderland, Massachusetts. Chapters contributed by many authors give the state-of-the-art in this exciting area.

Hartl, D. L., D. Freifelder and L. A. Snyder. 1988. *Basic Genetics*. Jones and Bartlett, Boston. A brief introduction to classical, population, and molecular genetics.

Hartl, D. L. 1988. *A Primer of Population Genetics*, 2nd Ed. Sinauer Associates, Sunderland, Massachusetts. An easily digested, thinner treatment than this book.

Hedrick, P. W. 1983. *Genetics of Populations*. Van Rostrand Reinhold, New York. Another text offering detailed coverage of classical population genetics.

Karlin, S. and E. Nevo. 1986. *Evolutionary Processes and Theory*. Academic Press, New York. Papers from a workshop on population genetics in Israel.

Kempthorne, O. 1957. *An Introduction to Genetic Statistics*. John Wiley and Sons, New York. Reprinted in 1969 by Iowa State University Press, Ames. This is a standard reference for the statistical tools needed for estimating parameters of quantitative genetics.

Kimura, M. 1983. *The Neutral Theory of Molecular Evolution*. Cambridge University Press, Cambridge. No student of population genetics should fail to read this book. A concise statement of an extremely influential theory.

Kingman, J. F. C. 1980. *Mathematics of Genetic Diversity*. Society for Industrial and Applied Mathematics, Philadelphia. A rigorous treatment of selected topics in population genetics theory. Especially good on the sampling theory of neutral genes and the coalescent process of the Wright-Fisher model.

Lewin, B. 1987. *Genes III*. John Wiley and Sons, New York. Population geneticists need to learn a lot of molecular biology, and this text provides a good background.

Lewontin, R. C. 1974. *The Genetic Basis of Evolutionary Change*. Columbia University Press, New York. This book remains essential reading for students of population genetics, providing a deep look at the nature of questions that can be addressed by population genetics.

Li, C. C. 1976. *First Course in Population Genetics*. Boxwood Press, Pacific Grove, California. The style of this book is rather different from any other, and appeals very much to some readers. Li likes to jump right in to numerical examples, and uses them very effectively.

Manly, B. F. J. 1985. *The Statistics of Natural Selection on Animal Populations*. Chapman and Hall, London. A thorough coverage of the subject, with a massive citation listing.

Mather, K. and J. L. Jinks. 1982. *Biometrical Genetics*, 3rd Ed. Chapman and Hall, London. The serious student of quantitative genetics should read both Falconer and this text. Mather and Jinks focus more on experimental designs for variance partitioning in plants.

Nagylaki, T. 1977. *Selection in One- and Two-Locus Systems*. Lecture Notes in Biomathematics Vol. 15. Springer-Verlag, Berlin. Hard core approach to the basic models.

Nei, M. and R. K. Koehn. 1983. *Evolution of Genes and Proteins*. Sinauer Associates, Sunderland, Massachusetts. Contains thirteen review papers by a variety of experts.

Nei, M. 1987. *Molecular Evolutionary Genetics*. Columbia University Press, New York. A rather comprehensive text on the subject, with emphasis on interpretation of molecular data.

Ohta, T. and K. Aoki. 1985. *Population Genetics and Molecular Evolution*. Springer-Verlag, New York. Collection of articles on various aspects of the neutral theory, including its history and role in population genetics.

Provine, W. B. 1971. *The Origins of Theoretical Population Genetics*. University of Chicago Press, Chicago. An engaging early history of theoretical population genetics.

Provine, W. B. 1986 *Sewall Wright and Evolutionary Biology*. University of Chicago

Press, Chicago. Superb scientific biography of one of the leading evolutionary biologists since Darwin.

Roughgarden, J. 1979. *Theory of Population Genetics and Evolutionary Ecology: An Introduction*. Macmillan, New York. This is the best text available on ecological genetics, and provides a clear derivation of diffusion approaches.

Weir, B. S., E. J. Eisen, M. M. Goodman and G. Namkoong (eds). 1988. *Proceedings of the Second International Conference on Quantitative Genetics*. Sinauer Associates, Sunderland, Massachusetts. Fifty-five contributed papers on all aspects of quantitative genetics, with a mix of reviews and original work.

Wright, S. *Evolution and the Genetics of Populations*. Vol. 1: *Genetic and Biometric Foundations* (1968). Vol. 2: *The Theory of Gene Frequencies* (1969). Vol. 3: *Experimental Results and Evolutionary Deductions* (1977). Vol. 4: *Variability within and among Natural Populations* (1978). University of Chicago Press, Chicago. A four-volume treatise by one of the founders of the field.

Answers to Chapter-End Problems

CHAPTER 1

1 The diploid can have n different homozygotes and $[n(n-1)]/2$ different heterozygotes. (The division by two is necessary because A_iA_j heterozygotes and A_jA_i heterozygotes are the same genotype.) The total number of possible genotypes is therefore $n + n(n-1)/2 = n(n+1)/2$. The tetraploid case is more complex because there are several kinds of heterozygotes. For the following genotypes the total possible numbers are

$A_iA_iA_iA_i$:	n
$A_iA_iA_iA_j$:	$n(n-1)$
$A_iA_iA_jA_j$:	$n(n-1)/2$
$A_iA_iA_jA_k$:	$n(n-1)(n-2)/2$
$A_iA_jA_kA_l$:	$n(n-1)(n-2)(n-3)/24$
Total	$(n+3)!/(n-1)!4!$

2 a. The allele frequency of $PGI\text{-}2^a = (70 + 19)/114 = 0.78$, while that of $PGI\text{-}2^b = 0.22$.

b. Expected genotype frequencies are $0.78^2 = 0.61$, $2(0.78)(0.22) = 0.34$, and $(0.22)^2 = 0.05$, giving expected numbers of 34.8, 19.4, and 2.8.

3 Allele frequency $q = (1/10{,}000)^{1/2} = 0.01$, and $2pq = 2(0.99)(0.01) = 0.0198$ is the frequency of carriers (i.e., about one person in 50).

4 The allele frequency of $d = (170/400)^{1/2} = 0.65$; of $D = 0.35$. Expected genotype frequencies of DD, Dd, and dd are 0.12, 0.46, and 0.42. Fraction of heterozygotes among Rh^+ is $0.46/(0.12 + 0.46) = 0.79$, and expected number of heterozygotes is $0.79 \times 230 = 182$.

5 The allele frequency of $M = (2202 + 1496)/6200 = 0.60$; of $N = 0.40$. Expected numbers are $(0.60)^2(3100) = 1116$, $2(0.60)(0.40)(3100) = 1488$, and $(0.40)^2(3100) = 496$. The $\chi^2 = (1101 - 1116)^2/1116 + (1496 - 1488)^2/1488 + (503 - 496)^2/496 = 0.34$ with one degree of freedom (because there are three classes of data and one parameter estimated from the data), with an associated probability value of about 0.65.

6 Expand $(0.1A_1 + 0.2A_2 + 0.3A_3 + 0.4A_4)^2$ to obtain A_1A_1: $(0.1)^2 = 0.01$; A_1A_2: $2(0.1)(0.2) = 0.04$; A_1A_3: $2(0.1)(0.3) = 0.06$; A_1A_4: $2(0.1)(0.4) = 0.08$; A_2A_2: $(0.2)^2 = 0.04$; A_2A_3: $2(0.2)(0.3) = 0.12$; A_2A_4: $2(0.2)(0.4) = 0.16$; A_3A_3: $(0.3)^2 = 0.09$; A_3A_4: $2(0.3)(0.4) = 0.24$; A_4A_4: $(0.4)^2 = 0.16$.

7 If the heterozygous parental genotype is *Aa*, half the gametes contain *A* and half contain *a*. The *A* gametes yields a heterozygous offspring when it unites with an *a* gamete (probability q) and the *a* gamete yields a heterozygous offspring when it unites with an *A* gamete (probability p). The overall probability of heterozygous offspring is $(1/2)q + (1/2)p = (1/2)(q + p) = 1/2$.

8 If the dominant and recessive allele frequencies are p and q, then $D = p^2$, $H = 2pq$, and $R = q^2$, so $DR = p^2q^2$ and $H = (2pq)^2$, confirming that $DR = H^2/4$.

9 Frequency of $AA = p^2 = (1 - q)^2 = 1 - 2q + q^2 \approx 1 - 2q$; frequency of *Aa* heterozygotes $= 2pq = 2(1 - q)q = 2q - 2q^2 \approx 2q$; and frequency of $aa = q^2 \approx 0$.

10 The probability that an individual with the dominant phenotype is heterozygous is $2pq/(p^2 + 2pq) = 2q/(1 + q)$, and among these individuals, the recessive allele frequency is 1/2; hence, the frequency of recessive alleles among dominant phenotypes is $[2q/(1 + q)](1/2) = q/(1 + q)$.

11 Allele frequency $q = 0.05$. The expected frequency of homozygous (color blind) females $= (0.05)^2 = 0.0025$ (about 1/400); the expected frequency of carriers $= 2(0.95)(0.05) = 0.095$ (about 1/10).

12 a. Using the formula from Box H (p. 40), the initial estimates are $\langle p \rangle = 0.169$, $\langle q \rangle = 0.106$, and $\langle r \rangle = 0.718$. These add up to 0.993, and when they are normalized by dividing each by 0.993, the estimates become 0.170, 0.107, and 0.723, respectively.

b. Applying the formula from Box H, the frequency of AA = freq(A) $(p^2 + 2pr) = (0.284)(0.029/0.275) = 0.030$. In a similar fashion, the frequencies of the remaining genotypes are: *AO*, 0.254; *BB*, 0.162; *OO*, 0.516; and *AB*, 0.026.

c. After the M step the revised allele frequencies become: $\langle p \rangle = [2(0.030) + 0.254 + 0.026]/2 = 0.170$; $\langle q \rangle = 0.106$; and $\langle r \rangle = 0.724$.

d. For phenotype A, the genotype could be *AA* or *AO*, with frequencies $p^2 + 2pr = 0.275$. The expected count is $(0.275)(155) = 42.650$. For the other phenotypes, the expected counts are: B, 25.496; O, 81.292; and AB, 5.302.

e. Calculating χ^2 from the observed and expected counts, we get $\chi^2 = 0.473$, which is not significant and indicates an excellent fit to Hardy-Weinberg proportions. There is 1 degree of freedom because there were 4 phenotypes − 2 estimated parameters − 1 = 1.

13 Allele frequency q = frequency of affected males. Frequency of carrier females = $2pq = 2(1 - q)q = 2q - 2q^2 \approx 2q$ when $q^2 \approx 0$. Therefore, the frequency of carrier females is approximately two times the frequency of affected males.

14 The genotype frequencies are given by the expansion of $(pA + qa)^4$ instead of $(pA + qa)^2$.

15 By the definition in the text, a polymorphic gene is one for which the most common allele has a frequency ≤ 0.95, making genes 1, 2, and 5 polymorphic in this sample, giving $P = 3/5 = 60\%$. Heterozygosity with random mating equals $1 - \Sigma p_i^2$, which for genes 1–5 equals 0.47, 0.11, 0.01, 0, and 0.37, with an average of $H = 0.19$.

16 There are $n = 10$ sequences, $m = 14$ cleavage sites, $k = 13$ polymorphic sites (all but site 11), $j = 6$ bases in the recognition sequence, and $c = \Sigma c_i = 56$ total number of cuts. Equation 1.1 estimates $\langle P_{\text{nuc}} \rangle = [56 - 10(14 - 13)]/(6)(56) = 0.14$, and Equation 1.2 uses $\Sigma c_i^2 = 378$ and estimates $\langle H_{\text{nuc}} \rangle = [(10)(56) - 378]/(6)(56)(10 - 1) = 0.06$.

17 The *normal* allele is dominant because of phenotype in F_1. The χ^2 value equals $(88 - 93.75)^2/93.75 + (37 - 31.25)^2/31.25 = 1.41$ with one degree of freedom (since no parameters were estimated from the data). The associated probability level is approximately 0.25, so the hypothesis of a 3:1 ratio cannot be rejected.

18 The expected ratio is obtained from the expansion of $[(3/4)D + (1/4)R]^3$, where D and R represent the dominant and recessive phenotypes, respectively, yielding the ratio 27:9:9:9:3:3:3:1 for each 64 progeny. Expectations are 269.6:89.9:89.9:89.9:30.0:30.0:30.0:10.0 among 639 progeny, for a χ^2 of 2.67 with 7 degrees of freedom (because there are 8 classes of data and no parameters estimated), for which the associated probability value is about 0.92. This is a very good fit, indeed.

19 Assuming no linkage, expected numbers are 50 and 50, so the χ^2 value is $(60 - 50)^2/50 + (40 - 50)^2/50 = 4.0$ with associated probability 0.048, which is statistically significant. Estimated recombination fraction is $40/100 = 0.40$.

20 **a.** Assuming a recombination fraction of 0.40, the expected numbers are 72, 48, 48, 72, for a χ^2 of 7.10 with three degrees of freedom (because the recombination fraction was not estimated from these data but from the independent experiment in Problem 19), for an associated probability level of 0.08, which is not statistically significant.

b. Altogether there were $40 + 76 = 116$ recombinants among $100 + 240 = 340$ progeny, for an estimated recombination fraction of $116/340 = 0.34$.

21 A_1B_1, $(0.3)(0.2) = 0.06$; A_1B_2, $(0.3)(0.3) = 0.09$; A_1B_3, $(0.3)(0.5) = 0.15$; A_2B_1, $(0.7)(0.2) = 0.14$; A_2B_2, $(0.7)(0.3) = 0.21$; A_2B_2, $(0.7)(0.5) = 0.35$.

22 a. Assuming linkage equilibrium, frequencies of gametes A_1B_1, A_1B_2, A_2B_1, A_2B_2 are p_1q_1, p_1q_2, p_2q_1, p_2q_2, or (0.7)(0.3) = 0.21, (0.7)(0.7) = 0.49, (0.3)(0.3) = 0.09, (0.3)(0.7) = 0.21, respectively.

b. The theoretical maximum of D equals the smaller of p_1q_2 (0.49) and p_2q_1 (0.09), or 0.09, and 50% of 0.09 gives $D = 0.045$. Gametic frequencies, given in the same order as in (a), are 0.21 + 0.045 = 0.255, 0.49 − 0.045 = 0.445, 0.09 − 0.045 = 0.045, 0.21 + 0.045 = 0.255.

23 When dominant × dominant matings occur at random, the proportion of homozygous recessive offspring equals the square of the recessive allele frequency among individuals with the dominant phenotype, or $[q/(1 + q)]^2$, because random mating of individuals is equivalent to random union of gametes, just as in the derivation of the Hardy-Weinberg principle. With dominant × recessive matings, the probability of a recessive gamete from the parent with the dominant phenotype is $q/(1 + q)$, and from the parent with the recessive phenotype it is 1; altogether the probability of a homozygous recessive offspring from a dominant × recessive mating is $[q/(1 + q)](1) = q/(1 + q)$.

24 One approach is to arrange the alleles in order:

$$A_1\ A_2\ A_3\ A_4\ \ldots\ A_n$$

and then insert symbols like | at random to the right of the alleles to represent the chromosomes. Any genotype will contain as many A_i alleles as there are | symbols to the immediate right of A_i. For example, with three alleles in a tetraploid, the string $A_1|A_2A_3|||$ represents the genotype $A_1A_3A_3A_3$. With these conventions, the number of possible ways to insert the p chromosomes symbols among the n alleles is

$$n(n + 1)(n + 2)\ \ldots\ (n + p - 1)$$

because the addition of each | symbol creates an additional possible space for the next one. However, there are $p!$ ways that the chromosome symbols can be permuted, so the total number of genotypes is

$$[n(n + 1)(n + 2)\ \ldots\ (n + p - 1)]/p!$$

Multiplying numerator and denominator by $(n - 1)!$, the expression takes the more compact form $(p + n - 1)!/[p!(n - 1)!]$.

CHAPTER 2

1 For an autosomal gene, the 24 cats represent 48 copies of the gene, so the probability of ultimate fixation is 1/48 and the probability of ultimate loss is 47/48. For an X-linked gene, the 24 cats represent 36 copies of the gene, so the probability of ultimate fixation is 1/36 and that of ultimate

loss 35/36. This is just an approximation for the X-linked case, because the effective size is not equal to the number of gene copies.

2 Equation 2.16 implies that $t = 1.39N$ generations will reduce the average heterozygosity by half. Because the plant is an annual, one generation corresponds to one year, so in this case $t = 50$, giving $N = 36$ as the effective population size.

3 Use Equation 2.10 with $N = 20$ and $t = 8$, which yields 0.18. Since self-fertilization cannot occur, use of $N = 20.5$ would be a bit more accurate, but this gives essentially the same answer. The value $F_{ST} = 0.18$ is not much less than the inbreeding coefficient of 0.25 resulting from one generation of brother–sister mating.

4 Using Equation 2.9,

$$t = \frac{\ln(1 - F_t) - \ln(1 - F_0)}{\ln[1 - (1/2N)]}$$

From $F = 0.01$ to 0.02 requires $N = 1$ generation; from 0.05 to 0.10 requires 5.4 generations. Taking logarithms of both sides of Equation 2.9 and using the suggested approximations yields $-F_t = -t/(2N) - F_0$, and when $F_t = 2F_0$, then $t = 2NF_0$. For $F_0 = 0.01$ or 0.05, the approximation gives $t = 1$ and $t = 5$ generations, respectively.

5 $N_f = 200$, and the number of breeding males equals $N_f/5 = 40$. In this situation $N_e = 4N_mN_f/(N_m + N_f) = 4(40)(200)/(40 + 200) = 133$.

6 Here $N_e = 4N_mN_f/(N_m + N_f)$. For 10 cows and one bull, $N_e = 4(1)(10)/(1 + 10) = 3.6$; for 40 cows and one bull, $N_e = 3.9$; for 10 cows and two bulls, $N_e = 6.7$.

7 The appropriate formula is Equation 2.28, which for $N_f = 100$ and $N_m = 10$ yields 37.5 and for $N_f = 10$ and $N_m = 100$ yields 21.4.

8 Use Equation 2.36 with $\delta = 15/m^2$ and $\sigma^2 = 0.5\ m^2$, which yields $N_e = 4(3.14)(15)(0.5) = 94$ individuals.

9 In the F_1 populationm, all restriction site differences are heterozygous so that the initial RFLP allele frequencies are 1/2. Equation 2.18 therefore applies with $N = 80$ and $L_0 = 100$. For $t = 10$, $L_t = 100\exp(-10/160) = 94$, and for $t = 50$, $L_t = 73$.

10 From Equation 2.25, $1/N_e = (1/n)[\Sigma(1/N_i)]$. In this case the $1/N_i$ are 1/2, 1/4, 1/8, . . . , $1/2^n$, which forms a geometric progression with $x = 1/2$, so the summation of the first n terms equals $(1/2)[1 - (1/2)^n]/(1/2) = 1 - (1/2)^n$. Therefore, $N_e = n/[1 - (1/2)^n]$ in the nth generation, and as n becomes large the limiting value of $N_e = n$.

11 Use Equation 2.35 with $V_k = 4$, which gives $N_e = (2/3)N - (1/3)$, or if N is not very small, approximately $(2/3)N$.

12 **a.** Since $V_k = 0$, $N_e = 2N - 1$.

b. Random genetic drift occurs in spite of the equal contributions because of Mendelian segregation.

c. In this case the effective population size is infinite; since the population is asexual and each individual contributes an exact replica of itself to the next generation, allele frequencies cannot change and random genetic drift does not occur.

13 Use Equation 2.38 with $N_e = 50$ and $n = 100$, which gives $T_{21}^n = 1 - (0.99)^{100} = 0.63$.

14 In a growing population, random individuals will share common ancestors more recently than predicted by the equation, so the equation is an overestimate of the time to coalescence.

15 In Equation 2.38, set $t = n = N$, implying $T_{21}^n = 1 - [1 - (1/2N)]^N \approx 1 - \exp[-(N)(1/2N)] = 1 - \exp(-0.5) = 0.39$. For $N = 50$, $T_{21}^n = 1 - (0.99)^{50} = 0.39$.

16 Use Equation 2.14 with $H_t = H_0/x$, which yields $1/x = \exp(-t/2N)$, or $\ln(1/x) = -t/2N$, which implies that $t = 2N\ln x$. For $x = 2$, the value $t = 1.39N$, agreeing with the value stated in the text.

17 Substitute $L_t = L_0/e$ into Equation 2.18 to obtain $\exp(-1) = \exp(-t/2N)$, whence $t = 2N$.

18 Average number of generations until fixation or loss are tabulated below.

Population	Average time to fixation	Average time to loss
1	320	8.5
2	640	9.6
3	1280	10.7
4	2560	11.8
5	5120	12.9
6	10,240	14.0
7	20,480	15.1
8	40,960	16.2
9	81,920	17.3
10	163,840	18.5

19 When $p = 0.5$, then $t_1 = -(2)[4N_e(1/2)\ln(1/2)] = 4N_e\ln(2)$, and $t_0 = -4N_e(1)\ln(1/2) = 4N_e\ln(2)$. The common value is about $2.8N_e$, which is the maximum average number of generations that a neutral gene persists in a population, whatever its initial frequency.

20 The transition matrix **T** consists of entries t_{ij} = Pr[progeny plant contains j copies of allele A, given parental plant contains i copies], hence

$$\mathbf{T} = \begin{pmatrix} 1 & 0 & 0 \\ 1/4 & 1/2 & 1/4 \\ 0 & 0 & 1 \end{pmatrix}$$

Then $\mathbf{x}_1 = \mathbf{x}_0\mathbf{T} = (0, 1, 0)\mathbf{T} = (1/4, 1/2, 1/4)$; $\mathbf{x}_2 = \mathbf{x}_1\mathbf{T} = (3/8, 1/4, 3/8)$; $\mathbf{x}_3 = (7/16, 1/8, 7/16)$. The general rule suggested for $\mathbf{x}_n$ is $[(2^n-1)/2^n, 1/2^{n-1}, (2^n-1)/2^n]$.

21 Again the transition matrix $\mathbf{T}$ consists of entries t_{ij} = Pr[progeny plant contains j copies of allele A, given parental plant contains i copies], but in this case $i, j = 1$ or 2. Since the ratio of Aa:AA among surviving plants from self-fertilization of an Aa plant equals 2/3:1/3,

$$\mathbf{T} = \begin{pmatrix} 2/3 & 1/3 \\ 0 & 1 \end{pmatrix}$$

Then $\mathbf{x}_1 = \mathbf{x}_0\mathbf{T} = (1, 0)\mathbf{T} = (2/3, 1/3)$; $\mathbf{x}_2 = \mathbf{x}_1\mathbf{T} = (4/9, 5/9)$; $\mathbf{x}_3 = (8/27, 19/27)$. The general rule suggested for $\mathbf{x}_n$ is $[(2/3)^n, 1 - (2/3)^n]$.

CHAPTER 3

1 Various kinds of mutations anywhere along the gene can result in loss of gene function; once mutated, only very specific kinds of reverse mutations will restore function.

2 Because the antibiotic-resistant cells in the colonies on the unselective plate were never themselves exposed to the antibiotic.

3 Four possible transitions and eight possible transversions for a ratio of 4:8 or (1/2):1. If observed ratio is 2:1, then, relative to their expected value, transitions are about 4 times too frequent.

4 $P_0 = 11/20 = 0.55$ and $\mu = -[\ln(P_0)]/N = 1.1 \times 10^{-9}$ per generation.

5 Use the Poisson zero term $P_0 = \exp(-m)$ with $P_0 = 1 - 0.35 = 0.65$, giving $m = -\ln P_0 = 0.43$ as the average number of lethals per chromosome.

6 For dominant lethals $(2 - 10 \times 10^{-2})/(5 \times 10^{-4}) = 40 - 200$ rads; for recessive visibles $(8 \times 10^{-6})/(7 \times 10^{-8}) = 114$ rads; for reciprocal translocations $(2 - 5 \times 10^{-4})/(1 - 2 \times 10^{-5}) = 10 - 50$ rads. (Incidentally, humans appear to be somewhat less sensitive to radiation than mice.)

7 With $m = 0$, $p_t = p_0(1 - \mu)^t$. For 10, 100, 1000, and 10,000 generations, p_t equals 0.99995, 0.9995, 0.995, and 0.95, respectively. Note that the approximation $p_t = 1 - \mu t$ is very accurate in this case.

8 Use $p_t = p_0(1 - \mu)^t$ with $p_0 = 1$, $\mu = 0.01$, and $p_t = 0.90$. Then $t = \ln(0.90)/\ln(0.99) = 10.5$ generations.

9 Use $q_t = q_0 + \mu t$. For $t = 0$ to 12 generations, each four-generation interval

increases q by 2×10^{-6}, so $\mu = (2 \times 10^{-6})/4 = 5 \times 10^{-7}$ per generation. (b) For $t = 12$ to 24, each interval increases q by 0.04×10^{-6}, so $\mu = (0.04 \times 10^{-6})/4 = 10^{-8}$ per generation. The novel metabolite decreases the mutation rate; such substances are called antimutagens.

10 **a.** Equilibrium $p = \mu/(\mu + \nu) = 1/11$
b. 1/101
c. 1/2
d. 1/11

11 Equation 3.2 implies that $(p_t - \hat{p})/(p_0 - \hat{p}) = (1 - \mu - \nu)^t$, and halfway to equilibrium means that $p_t - \hat{p} = (1/2)(p_0 - \hat{p})$, so $(1 - \mu - \nu)^t$ equals 1/2. Therefore $t \ln(1 - \mu - \nu) = \ln(1/2)$ or, using the approximation, $t \approx \ln(1/2)/(\mu + \nu) \approx 0.7/(\mu + \nu)$. For the values of μ and ν specified, $t \approx 6.4 \times 10^4$ generations.

12 $q_1 = q_0 + \mu_0$; $q_2 = q_1 + \mu_1 = q_0 + \mu_o + \mu_1$; and so on to obtain $q_t = q_0 + \Sigma\mu_i$. This can be written as $q_0 + \mu t$ if μ is interpreted as the arithmetic mean of the mutation rates $(\Sigma\mu_i)/t$.

13 Use $F_t = 1 - [1 - 1/(2N)]^t$ with $N = 50$ and $t = 200$, so $F_{200} = 0.866$.

14 Use $F_t = 1 - [1 - 1/(2N)]^t \approx 1 - \exp(-t/2N)$ with $F_t = 1/4$. Thus $t \approx -2N\ln(3/4) \approx 0.6N$ generations.

15 Effective number of alleles = reciprocal of homozygosity (sums of squares of allele frequencies). These examples have $n_e = 2.8$ and 8.

16 Because rare alleles contribute negligibly to the homozygosity.

17 Use Equation 3.20, $E(r) = 1 + \theta/(\theta + 1) + \theta/(\theta + 2) + \ldots + \theta/(\theta + 9)$. For $\theta = 0.1$, 1, and 10, $E(r) = 1.27$, 2.93, and 7.19, respectively.

18 Use Equation 3.13 with $N = 50$, $\mu = 10^{-5}$ and $m = 10^{-3}$ to obtain $\hat{F} = 0.83$, so $\hat{H} = 1 - \hat{F} = 0.17$.

19 Use Equation 3.20 with $E(r) = 1 + \theta/(\theta + 1)$ and $\theta = x/(1 - x)$, which yields $1 + x$.

20 Problem 19 notes that the expected number of neutral alleles in a sample of size two equals $1 + x$ when $\theta = x/(1 + x)$. Since $H = \theta/(1 + \theta)$ (Equation 3.15), the heterozygosity is $H = x$.

21 The expected value of Equation 3.29 is $4N/n(n - 1)$. When $n = \sqrt{N}$, the expected value is $4N/\sqrt{N}(\sqrt{N} - 1) \approx 4$ generations for large N. For $N = 60$ and $n = 8$ the answer is 4.3 generations.

CHAPTER 4

1 Because homozygotes for some alleles may be superior to heterozygotes containing other alleles.

2 Fitnesses are A_1A_2 0.7; A_1A_3 0.6; A_1A_4 0.5; A_2A_3 0.5; A_2A_4 0.4; A_3A_4 0.3; and $p_i = 1/4$; $\overline{w} = [\Sigma(A_iA_i \text{ fitnesses}) + 2\Sigma(A_iA_j \text{ fitnesses})]/16 = 0.5$.

3 Use Equation 4.3, giving $p_1 = 0.736$, $p_2 = 0.768$, $p_3 = 0.796$.

4 Equilibrium $\hat{p} = t/(s + t)$ the fitnesses are $1 - s$, 1, and $1 - t$.

a. $s = 0.7$, $t = 0.3$

b. $s = 0.07$, $t = 0.03$

c. $s = 0.007$, $t = 0.003$. In all cases $\hat{p} = 0.3$.

5 $\overline{w} = (0.64)(0.9) + 2(0.8)(0.2)(1) + (0.04)(0.6) = 0.92$. Note that $0.8 = 0.4/(0.4 + 0.1)$, so 0.8 is the equilibrium frequency for this case of overdominance, and in this model the equilibrium occurs at maximum $\overline{w}$.

6 $q' = pq(1 - h)/[p^2 + 2pq(1 - h)] = q(1 - h)/(1 - qh) \approx q(1 - h)$ with $q'/q = 0.99$ given, so $h = 0.01$.

7 From Equation 4.63, $\hat{q} = \mu/hs$ with $hs = 1/2$, hence $\hat{q} = 4.5 \times 10^{-5}$; expected frequency of affected individuals is $2\hat{p}\hat{q} \approx 9 \times 10^{-5}$.

8 From Equation 4.62, $\hat{q} = \sqrt{(4 \times 10^{-6})/0.2)} = 4.5 \times 10^{-3}$; from Equation 4.63, $\hat{q} = 4 \times 10^{-6}/(0.05)(0.2) = 4.0 \times 10^{-4}$. The 5% heterozygous effect reduces the equilibrium frequency by more than an order of magnitude.

9 From Equation 4.62, $\hat{q} = \sqrt{(\mu/s)} = \sqrt{(10^{-6}/0.6)} = 1.3 \times 10^{-3}$; would be 1.0×10^{-3} if homozygotes did not reproduce (in which case $s = 1$).

10 Frequency of heterozygotes $\approx 2\hat{q}$, frequency of heterozygotes from new mutations $\approx 2\mu$, ratio $= \mu/\hat{q}$, and since $\hat{q} = \mu/h$, ratio $= h$.

11 $t = 4[-2.2 + 11.5] = 37.2$ generations. Note that the equation can be written as $t = (2/s)[\ln(p_t/p_0) - \ln(q_t/q_0)]$, and with p_t, $p_0 << 1$, $\ln(q_t/q_0) \approx 0$. For a large range of values of p_t/p_0 and s, t ranges from 5 to 50. This explains in part why significant pesticide resistance evolves so rapidly in diverse species.

12 For the haploid, $p' = p/(1 - qs)$. For the diploid, $p' = [p^2 + pq(1 - s)]/[p^2 + 2pq(1 - s) + q^2(1 - s)^2] = p[p + q(1 - s)]/[p + q(1 - s)]^2 = p/[p + q(1 - s)] = p/(1 - qs)$.

13 $p_1 = p_0s_0/(p_0s_0 + q_0)$; $q_1 = q_0/(p_0s_0 + q_0)$; and so on gives $(p_1/q_1) = (p_0/q_0)s_0$; $p_2/q_2 = (p_0/q_0)(s_0s_1)$; . . . ; $p_n/q_n = (p_0/q_0)(s_0s_1 \ldots s_{n-1})$. To use the formula with s, set $s = (s_0s_1 \ldots s_{n-1})^{1/n}$, so s is the geometric mean of the selection coefficients.

14 Interchange p and q and change sign of s to obtain $\ln(q_t/p_t) + 1/pt = [\ln(q_0/p_0) + 1/p_0] - st$, or $\ln(p_t/q_t) - 1/p_t = [\ln(p_0/q_0)] - 1/p_0] + st$, where p_t is now the allele frequency of the favored recessive and s the selection coefficient against individuals with the dominant allele.

15 $d\Delta p/dp = (1/2 - p)(1 - p) - p(1/2 - p) - p(1 - p)$. This is positive for $p = 0$ and $p = 1$ (indicating unstable equilibria) and negative for $p = 1/2$ (indicating local stability). The latter point is also globally stable.

16 Condition 4.31 requires $(v_{12}/0.5) + (1/0.8) > 2$, or $v_{12} > 0.375$; condition 4.32 requires $(1/0.9) + (v_{12}/1) > 2$, or $v_{12} > 0.89$. The latter, being the more restrictive, is the required condition for protection.

17 Use Equation 4.70 with $w_{11} = 1$, $w_{12} = 1$, $w_{22} = 0$, $v_1 = 1 - s$, $v_2 = 1$, so $\hat{p} = 1 - s$ is the frequency of A, giving $\hat{q} = s$ as the frequency of a.

18 Set $\hat{p} = 1 - 0.125 = 0.875$, $w_{11} = 1$, $w_{22} = 0$, and $k = 0.75$. Then w_{12} satisfies $-2(0.25)w_{12}/(1 - 2w_{12}) = 0.875$, or $w_{12} = 0.7$. Actually, *segregation distorter* is active only in males, and a model that takes this into account yields $w_{12} = 0.79$ in males and $w_{12} = 1$ in females (Hartl 1970).

19 For a recessive lethal, $w_{11} = w_{12} = 1$, $w_{22} = 0$. Then $q' = pq/(p^2 + pq) = q/(1 + q)$, or $1/q_n = 1 + 1/q_{n-1}$; which implies $1/q_n = n + 1/q_0$, or $q_n = q_0/(1 + nq_0)$. For $q_n = q_0/2$, $n = 1/q_0$ generations.

20

p_0	$N_s = 0.1$	$N_s = 10$
0.1	0.11	0.86
0.2	0.22	0.98
0.3	0.32	1.00
0.4	0.42	1.00
0.5	0.52	1.00
0.6	0.62	1.00
0.7	0.72	1.00
0.8	0.82	1.00
0.9	0.91	1.00

21 Equation 4.74 assumes fitnesses $1 + s'$, $1 + s'/2$ and 1, which when compared to $1 + 2s$, $1 + s$ and 1 gives $s = s'/2$. Substituting this into the alternative equation gives $\text{Prob}(A \text{ fixes}) = 2(s'/2)/\{1 - \exp[-2N(s'/2)]\} = s'/[1 - \exp(-2Ns')]$, which is Equation 4.74. The only difference is in how the fitnesses are scaled.

22 **a.** $\alpha = (pw_{11} + qw_{12}) - (pw_{12} + qw_{22}) = w_1 - w_2$.

b. $pw_1 + qw_2 = p^2w_{11} + pqw_{12} + pqw_{12} + q^2w_{22} = \overline{w}$.

c. $w_1 - \overline{w} = w_1 - (pw_1 + qw_2) = q(w_1 - w_2) = q\alpha$ and $w_2 - \overline{w} = -p\alpha$, so $V_g = 2pq^2\alpha^2 + 2p^2q\alpha^2 = 2pq\alpha^2$.

d. $\Delta p = p[q\alpha]/\overline{w} = pq\alpha/\overline{w}$.

23 $\overline{w}' = \overline{w} + 2w_{11}p\Delta p + 2w_{12}q\Delta p - 2w_{12}p\Delta p - 2w_{22}q\Delta p + (\Delta p)^2(w_{11} - 2w_{12} + w_{22}) = \overline{w} + 2\Delta p(w_1 - w_2) + (\Delta p)^2 (w_{11} - 2w_{12} + w_{22}) = \overline{w} + 2pq\alpha^2/\overline{w} + (p^2q^2\alpha^2/\overline{w}^2)(w_{11} - 2w_{12} + w_{22}) = \overline{w} + (V_g/\overline{w})[1 + (V_g/2\overline{w})(w_{11} - 2w_{12} + w_{22})]$.

CHAPTER 5

1 Use the same algorithm as for X-linked genes in otherwise diploid organisms.

2 Because a male contributes the Y chromosome to his sons, paths with two or more consecutive males have probability 0 for the transmission of an X-linked gene.

3 $F = 1/2$ in odd-numbered generations (the progeny of selfing), and $F = 0$ in even-numbered generations (the progeny of random mating).

4 $F_I = (1/2)^2(1 + F_A) = 5/16$.

5 $F_I = (1/2)(1 + F_A)$ and $F_A = (1/2)^3$, so $F_I = 9/16$.

6 $p^2 - p^2F + pF = p^2 + p(1 - p)F = p^2 + pqF = p(1 - q) + pqF = p - pq + pqF = p - pq(1 - F)$.

7 Random mating frequencies are 1/4, 1/2, and 1/4; among the offspring of first cousins ($F = 1/16$), they are 17/64, 30/64, 17/64. Deficiency of heterozygotes $= (32/64 - 30/64)/(32/64) = 1/16 = F$.

8 $q = (1/1600)^{1/2} = 1/40$, and expected frequency from first-cousin matings is $(1/1600)(15/16) + (1/40)(1/16) \approx 1/465$.

9 Proportions with first-cousin parents are 0.016, 0.022, 0.068, 0.120, and 0.391 for the frequencies given. When $q = 1$, the proportion is 0.01, which says that 1% of the matings are between first cousins.

10 Paths GEF, GEBCF, GDBCF, GDACF, GDBEF, so $F_I = (1/2)^3 + 4(1/2)^5 = 1/4$.

11 A_1A_2, A_2A_3, A_1A_3 styles in equal frequency, so probability of any pollen grain landing on a compatible style is 1/3.

12 Equilibrium $F = 0.2/(2 - 0.2) = 1/9$. Genotype frequencies are $(1/9)(8/9) + (1/3)(1/9)$, $(4/9)(8/9)$, $(4/9)(8/9) + (2/3)(1/9)$, or 0.136, 0.395, 0.469.

13 Coefficient of relationship is the inbreeding coefficient of the hypothetical offspring K, so it equals $4(1/2)^7 = 1/32$.

14 $F_A = F_B = 0$; $F_C = F_D = 0$; $F_E = F_F = 2(1/2)^3 = 1/4$; $F_G = F_H = 2(1/2)^3 + 4(1/2)^5 = 3/8$; $F_I = (1/2)^3(1 + F_E) + (1/2)^3(1 + F_F) + 4(1/2)^5 + 8(1/2)^7 = 8/16$. Note that the pedigree is of three generations of sib mating.

15 $F_0 = 0$, $F_1 = 0$, $F_2 = 1/4$, $F_3 = 3/8$, $F_4 = 8/16$, $F_5 = 19/32 = 0.59$, so genotype frequencies after 5 generations are 0.21, 0.17, and 0.61; one additional generation of random mating restores the Hardy-Weinberg frequencies of 0.09, 0.42, and 0.49.

16 Rewrite Equation 5.19 as $F_{t+1} - C = (S/2)(F_t - C) = (S/2)^2(F_{t-1} - C) = \ldots = (S/2)^{t+1}(F_0 - C)$. With $S = 1/2$ and $F_0 = 0$, $F_1 = 0.25$, $F_2 = 0.31$, $F_3 = 0.33$. The equilibrium $\hat{F} = C = 1/3$.

17 $x_i' = [1/(n - 1)](1 - x_i)$ so $\hat{x}_i = 1/n$.

18 Total frequency of homozygous recessives equals $(1 - c)q^2 + c[q^2(15/16) + q(1/16)]$, where the second term comes from first-cousin matings. Ratio of second term to total simplifies to formula given.

19 Two-way hybrid has genotype A_1A_2 and gametes 1/2 A_1 + 1/2 A_2, so probability of identity by descent (F) among offspring is $(1/2)^2 + (1/2)^2 = 1/2$; three-way hybrid has gametes 1/4 A_1, 1/4 A_2, 1/2 A_3, and $F = (1/4)^2 + (1/4)^2 + (1/2)^2 = 3/8$; four-way hybrid has gametes 1/4 A_1, 1/4 A_2, 1/4 A_3, 1/4 A_4, and $F = 4(1/4)^2 = 1/4$.

20 $F_t = (1/2)^2(1 + F_{t-2}) + (1/2)^3(1 + F_{t-3}) + (1/2)^4(1 + F_{t-4}) + \ldots = (1/2)^2(1 + F_{t-2}) + (1/2)F_{t-1} = 1/4 + (1/2)F_{t-1} + (1/4)F_{t-2}$. Therefore, $F_0 = 0$, $F_1 = 0$, $F_2 = 1/4$, $F_3 = 3/8$, $F_4 = 8/16$, $F_5 = 19/32$.

21 $F_t = (1/2)^2(1) + (1/2)^3(1) + (1/2)^4(1) + \ldots = (1/2)^2 + (1/2)F_{t-1}$; therefore $F_0 = 0$, $F_1 = 1/4$, $F_2 = 3/8$, $F_3 = 7/16$, $F_4 = 15/32$, $F_5 = 31/64$. The equilibrium is given by $\hat{F} = (1/4) + \hat{F}/2$, or $\hat{F} = 1/2$.

22 $F_t = (1/2)^2(1 + F_A) + (1/2)^3(1 + F_A) + (1/2)^4(1 + F_A) + \ldots = (1/4)(1 + F_A) + F_{t-1}/2$. Setting $\hat{F} = F_t = F_{t-1}$ gives $\hat{F} = (1/2)(1 + F_A)$.

CHAPTER 6

1 $m = 1$.

2 Use Equation 6.17 with $p_t = 0.5$, $p_0 = 0.2$, $\bar{p} = 0.8$ and $m = 0.01$. Then $t = 69$ generations.

3 $(1 - m)^{10} = 0.6$, so the allele frequencies are 0.32, 0.44, 0.56, and 0.68.

4 The average frequency before fusion $= [(q + \epsilon)^2 + (q - \epsilon)^2]/2 = q^2 + \epsilon^2$; after fusion $= q^2$; difference $= \epsilon^2$ = variance in allele frequency among subpopulations.

5 q_1 in population 1, q_2 in population 2, $(q_1 + q_2)/2$ in fused population. There is no Wahlund effect in this situation because of haploidy of the X in males. There is, however, a Wahlund effect for X-linked genes in females.

6 Multiply $(1 - F_{IS})(1 - F_{ST}) = 1 - F_{IT}$ and cancel the 1's. The expression says that the probability of autozygosity within the total population equals the probability of autozygosity due to inbreeding within a subpopulation, plus the probability of autozygosity due to random genetic drift, minus the probability of autozygosity for both reasons.

7 Equation 6.17 is $p_t - \bar{p} = (p_0 - \bar{p})(1 - m)^t$, so $\Sigma(p_t - \bar{p})^2 = \Sigma(p_0 - \bar{p})^2(1 - m)^{2t}$, or $\sigma_t^2 = \sigma_0^2 (1 - m)^{2t}$.

8 Gametic frequencies are $p_1q_1 + \epsilon$, $p_1q_2 - \epsilon$, $p_2q_1 - \epsilon$, $p_2q_2 + \epsilon$ in one population, $p_1q_1 - \epsilon$, $p_1q_2 + \epsilon$, $p_2q_1 + \epsilon$, $p_2q_2 - \epsilon$ in the other; D in fused population equals 0. Notice that, although there is no gametic linkage disequilibrium, the genotype frequencies in the fused population are not in Hardy-Weinberg equilibrium.

9 Heterozygosities are 0.54, 0.62, 0.66, respectively, average 0.61. Fused population has allele frequencies 0.2, 0.3, 0.5, and heterozygosity 0.62. $F_{ST} = 0.02$. Maximum F_{ST} occurs when each population is fixed for a different allele, and $F_{ST} = 1$ for the set of three as well as for each pairwise comparision.

10 $F_t = 1 - [1 - 1/(2N)]^t$ where $F_t = F_{ST}$ and $F_0 = 0$. Therefore, $\sigma^2 = \bar{p}\bar{q}[1 - (1 - 1/(2N)]^t \approx \bar{p}\bar{q}[1 - \exp(-t/2N)]$.

11 Use Equation 6.24 with $1/(4Nm + 1) < 0.05$, so $m > 4.75/N$.

12 Allele frequency 0.2 in both subpopulations, giving $H_S = 0.32$. Average $H_I = 0.272$. $F_{IS} = 0.15$. (Note from the genotype frequencies that the populations have inbreeding coefficients 0.1 and 0.2, respectively, which average 0.15). Since the allele frequencies are identical, $H_T = 0.32$ also, so $F_{ST} = 0$. Since $(1 - F_{IT}) = (1 - F_{IS})(1 - F_{ST})$, $F_{IT} = 0.15$.

13 With inbreeding F, mean allele frequency equals $[p^2(1 - F) + pF](1) + [2pq(1 - F)](1/2) = p$. Mean square MS $= [p^2(1 - F) + pF](1) + [2pq(1 - F)](1/2)^2$. Variance equals MS $- p^2 = (pq/2)(1 + F)$. When $F = 0$ (random mating), variance $= pq/2$. When $F = 1$ (complete inbreeding), variance $= pq$.

14 With self-fertilization and $F_0 = 0$, $F_n = 1 - (1/2)^n$. Using the formula from Problem 13, the variance in generation n is $(pq/2)[2 - (1/2)^n]$.

15 From selection, $\Delta q = -q^2/(1 + q)$; and for migration, $\Delta q = m$. Therefore, at equilibrium, $q^2/(1 + q) = m$, or since m is small, $q \approx \sqrt{m}$. This is the same expression as obtained when new recessive lethal alleles arise by mutation.

16 **a.** $F_{ST} = 1 - (1 - 1/100)^{50} = 0.39$; $F_{IS} = 1/4$; $F_{IT} = 1 - (0.75)(0.61) = 0.54$. (The exponent in F_{ST} is 50, not 47, because random drift still occurs during the generations of sib mating.)

b. $F_{ST} = 0.39$; $F_{IS} = 0$; $F_{IT} = 0.39$.

17 $p_1 - \bar{p} = (p_0 - \bar{p})(1 - m_0)$; $p_2 - \bar{p} = (p_1 - \bar{p})(1 - m_1)(1 - m_0)$, and $p_t - \bar{p} = (p_0 - \bar{p})(1 - m_{t-1})(1 - m_{t-2}) \ldots (1 - m_0)$. These can be written as $p_t - \bar{p} = (p_0 - \bar{p})(1 - m)^t$ if $1 - m$ is interpreted as the geometric mean of the $1 - m_i$ values.

18 $H_1 = 0.62$, $H_2 = 0.48$, $\overline{H}_S = 0.55$, $H_T = 0.615$, $F_{ST} = 0.106$. $J_S = [(0.04 + 0.36)/2 + (0.09 + 0)/2 + (0.25 + 0.16)/2] = 0.45$, $J_T = [(0.4)^2 + (0.15)^2 + (0.45)^2] = 0.385$, $G_{ST} = 0.106$. Note that $J_S = 1 - \overline{H}_S$; $J_T = 1 - H_T$; so $G_{ST} = F_{ST}$.

19 $F_{ST(1)} = (0.48 - 0.40)/0.48 = 0.167$; $F_{ST(2)} = (0.255 - 0.21)/0.255 = 0.176$; $F_{ST(3)} = (0.495 - 0.49)/0.495 = 0.010$. Average frequencies are $\bar{p}_1 = 0.4$, $\bar{p}_2 = 0.15$, $\bar{p}_3 = 0.45$, and weighted average $F_{ST} = 0.106$, which equals the G_{ST} calculated in Problem 18.

20 $J_{11} = 0.30$, $J_{22} = 0.52$, $J_{12} = 0.32$. $J = 0.32/[(0.38)(0.52)]^{1/2} = 0.72$, $D = -\ln(I) = 0.33$.

CHAPTER 7

1 If you assume that most differences are synonymous (third position), then reading frame begins with position 1 at the far left. The sequences code for 20 amino acids, of which 2 are different. Altogether there are 60 nucleotides (13 different), 40 nonsynonymous sites (2 different), 20 synonymous sites (11 different).
   a. Use Equation 7.18 with $D = 2/20$, $\langle k \rangle = 0.105$.
   b. Use Equation 7.21 with $D = 13/60$, $\langle k \rangle = 0.256$.
   c. Use Equation 7.21 with $D = 2/40$, $\langle k \rangle = 0.052$.
   d. Use Equation 7.21 with $D = 11/20$, $\langle k \rangle = 0.991$.

2 With $\langle D \rangle = 0.33$ in Equation 7.21, $\langle k \rangle = 0.43$, and $t = 0.43/(2 \times 0.01/\text{yr}) = 21.5$ yr. The date of divergence is approximately 1983 − 21.5, or 1961–1962.

3 Values of $\langle k \rangle$ from Equation 7.21 are

	HIV2	VISNA	MMLV
HIV1	0.45	0.95	1.31
HIV2		0.89	1.37
VISNA			1.37

Data indicate that HIV1 and HIV2 most closely related to each other, more distantly to VISNA, and still more distantly to MMLV. HIV1, HIV2, and VISNA are all about equally distant from VISNA.

4 The selective constraints are probably weak, as 5×10^{-9} approximates the rate of synonymous substitution in many genes of many organisms.

5 Most likely the protein is undergoing very rapid amino acid replacement because of natural selection.

6 Compensatory substitutions, i.e., an A → G substitution at a site matched with a T → C substitution at the position with which it pairs. This is the pattern actually observed.

7 $D = 3/4$, which means that the nucleotide differences are as great as expected in comparing two random sequences.

8 $0.005 \times 10^{-6} = 5 \times 10^{-9}$ substitutions along a lineage. The rate of divergence between two lineages is twice this value, or 1% per million years.

9

i	$\Theta = 1/4$	$\Theta = 3/4$
0	0.25	0.25
1	0.56	0.19
2	0.14	0.14
3	0.035	0.11
4	0.0088	0.079
≥5	0.0062	0.231

10 Expected numbers are 14.2, 9.4, 7.9, 6.6, 5.5, and 27.4, with $\chi^2 = 3.6$.

This has 3 degrees of freedom (because α and Θ were estimated from the data), and $P = 0.34$.

11 Rate of evolution = $(3.6/2) \times 13 \times 10^6 = 1.39 \times 10^{-7}\ \Delta T_m$/yr, or ΔT_m = 1°C per 7.2×10^6 yrs. Given ΔT_m of 1°C = 1% mismatches, rate of nucleotide substitution is $0.01/(7.2 \times 10^6) = 1.4 \times 10^{-9}$/site/yr.

12 **a.** $3N/4$.
b. $N/4$.
c. $N/4$.

13 Nucleotide sites examined in each individual = $17 \times 6 = 102$. The proportion of polymorphic nucleotide sites is estimated as $\langle p \rangle = 1 - S^{1/6}$, where S is estimated from Equation 7.12. Then $\langle P \rangle = 1 - (78/89)^{(1/6)} = 0.022$.

14 $M = \exp(-2rd)$ with $r = 6$ and $d = 0.01$, giving $M = 0.887$. Then $F = 0.50$ is the fraction of shared restriction fragments.

15 F in Equation 7.9 equals $66/92 = 0.717$, and iteration of Equation 7.10 with $M = (0.717)^{1/4} = 0.92$ produces, after 5 iterations, $M = 0.945$. Then Equation 7.11 implies $\langle d \rangle = 0.038$, since $r = 6$ in this case.

16 **a.** Set $4N = 10 \times 2\ln(2N)$ and iterate. Convergence is very rapid for any starting value, and $N = 17.9$;
b. $N = 323.6$.

17 Set $f = c_2$ and solve for $\lambda = (n - 1)/4N$. Then $c_1 = c_2 = 1/(1 + 4N\mu] \approx 1$ because $4N\mu << 1$.

18 $c_1 = c_2 = 1/n$; $f \approx 1$.

19 When t = half life, $e^{-ht} = 1/2$, so $t = -\ln(1/2)/h$.

20 Distribution of persistence time = $h \times \exp(-ht)$ because

$$\int_0^t h\exp(-ht)\, dt = 1$$

Average persistence time is

$$\int_0^t ht \exp(-ht)\, dt = 1/h$$

CHAPTER 8

1 Directional dominance (i.e., genes primarily in plus or minus direction). Natural selection acting more strongly to resist progress in plus or minus direction.

2 $\langle \mu \rangle = 2206/30 = 73.53$, $\langle \sigma^2 \rangle = [(166{,}206/30) - (2206/30)^2](30/29) = 137.64$, $\langle \sigma \rangle = 11.73$.

3 **a.** 18–22 is one standard deviation, so a proportion $1 - 0.317 = 0.683$.
b. >22 is greater than one standard deviation, so proportion = $0.317/2 = 0.158$.

c. >24 proportion = 0.046/2 = 0.023.

d. 20–22 proportion = 0.5 − 0.158 = 0.342.

e. <16 proportion = 0.046/2 = 0.023.

4 $\sigma_e^2 = 1.5$; $\sigma_g^2 = 6.0 - 1.5 = 4.5$; broad sense heritability = 4.5/6.0 = 75%.

5 The phenotypic standard deviation is 200 μg, so $S = 400$ μg; $h^2 = 10{,}000/40{,}000 = 25\%$, so expected mean = 2000 + (0.25)(400) = 2100.

6 After one generation, $\mu' = 20 + 0.25(4) = 21$; after 10 generations (using Equation 8.38), $\mu' = 20 + 0.25(4)(10) = 30$ bristles.

7 h^2 = response/cumulative selection differential = 0.15/(0.07 × 5) = 0.43.

8 **a.** $a = (23.8 - 19.4)/2 = 2.2$; $d = [2(25.2) - 23.8 - 19.4]/2 = 3.6$.

b. Write values as if they were relative fitnesses, i.e., 1 − 0.056, 1, 1 − 0.230 for AA, AA', $A'A'$. Since maximum of mean fitness occurs at $p = 0.230/(0.230 + 0.056) = 0.804$, this value will maximize the mean of the quantitative trait.

9 Using Equation 8.40, $D = 3.73 + 2.01 = 5.74$ and $\sigma_g^2 = (0.87)^2 - (0.60)^2 = 0.40$, hence $\langle n \rangle = (5.74)^2/8(0.40) = 10.3$.

10 Let X = first litter, while Y = second litter. Then $\Sigma X = 104$, $\Sigma X^2 = 1106$, $\Sigma Y = 103$, $\Sigma Y^2 = 1101$, $\Sigma XY = 1089$, $r = [1089 - (104)(103)/10]/\{[106 - (104)^2/10][1101 - (103)^2/10]\}^{1/2} = 0.57$.

11 Response/$h^2 = (220 - 180)/0.20 = 200$ = Cumulative selection differential. Number of generations = 200/20 = 10.

12 Selection intensity = $i = [1/\sqrt{(2\pi)}]\exp(-t^2/2)/B$, where, taking the top half for example, $B = 1/2$ and $t = 0.0$, so $i = 0.80$. For the other proportions (in order), $i = 1.27, 1.63, 1.95, 2.26$.

13 Use subscripts M and F for males and females. $S_M = \mu - \mu_M$, $S_F = \mu - \mu_F$. Mean of selected parents is $(\mu_M + \mu_F)/2$, so proper total selection differential is $\mu - (\mu_M + \mu_F)/2 = (S_M + S_F)/2$.

14 For A and B loci, $2pq[a + (q - p)d]^2 = 0.18$ and 0.067, respectively, so $\sigma_a^2 = 0.247$, and $h^2 = \sigma_a^2/\sigma_p^2 = 0.247$. For A and B, $\sigma_d^2 = (2pqd)^2 = 0.01$ and 0, so broad sense heritability = 0.257.

15 Equation 8.10 can be written as $R = i\sigma h^2$ where i is the intensity of selection. For percent protein, $R = (1.5)(0.45)(0.7) = 0.47$, so expected percent protein = 3.3% + 0.47% = 3.8%. For correlated response use Equation 8.45, $CR = (1.5)(0.60)^{1/2}(0.70)^{1/2}(0.55)(0.65) = 0.35$, so expected percent fat = 3.4% + 0.35% = 3.75%. The fat increase corresponds to a direct selection of $i = 0.35/(0.60)(0.65) = 0.90$.

16 Let the phenotypes of AA, AA', $A'A'$ be 1, 0, 0 with frequencies p^2, $2pq$, q^2, so A' is dominant with frequency q. Then $a = 1/2$, $d = -1/2$, and $\sigma_a^2 = 2p^3q$. Then $\sigma_p^2 = p^2 - (p^2)^2 = p^2q(1 + p)$. Therefore $h^2 = 2p^3q/p^2q(1 + p) = 2(1 - q)/(2 - q)$, which is approximately $1 - q$ when q is small.

17 Let fitnesses of AA, AA', $A'A'$ be $1 - s$, 1, $1 - t$, so $a = -(s - t)/2$ and $d = (s + t)/2$. At equilibrium $p = t/(s + t)$ and $q = s/(s + t)$, so $a + (q - p)d = 0$ and $\sigma_a^2 = 0$. Note that this result is consistent with the fundamental theorem of natural selection.

18 For $B = 1/2$, 1/4, 1/8, 1/16, and 1/32, approximation gives $i = 0.80$, 1.25, 1.60, 1.91, and 2.21, respectively.

19 $Z = [1/\sigma\sqrt{(2\pi)}]\exp[-(T - \mu)^2/2\sigma^2]$ and $z = [1/\sqrt{(2\pi)}]\exp(-t^2/2) = [1/\sqrt{(2\pi)}]\exp[-(T - \mu)^2/2\sigma^2]$. Therefore, $z = \sigma Z$.

20 Since $dN(x)/dx = -xN(x)$, then $\int_t^\infty xN(x)dx = -N(x) = N(t) - N(\infty) = N(t) = z$. By definition, $\mu_S = \int_t^\infty xN(x)dx/\int_t^\infty N(x)dx = z/B$. Equation 8.11 is the more general case.

CHAPTER 9

1 The dominant case is more rapidly selected; the recessive case is more rapidly reversed.

2 Selection for a high intrinsic rate of increase because the repeated harsh conditions would be expected to keep the population size well below the carrying capacity.

3 Because founder events that can establish clines are not likely to be reproduced on different continents.

4 Because it is difficult to accumulate the needed genetic divergence in sympatric populations, whereas with geographic isolation, after a sufficient period of time, speciation becomes very likely.

5 Because the recipients of the altruism share genes with the altruists, and the cost–benefit ratio $c/b < 2F_{XY}/(1 + F_X)$ is easier to satisfy.

6 Because such treatment imposes a constant selection pressure for antibiotic resistance. This can be particularly serious in a hospital environment, where multiply-resistant bacterial strains have been isolated. Although the immune system also imposes selection pressure on bacteria, it can adapt with antigenic changes.

7 (1) The ability to be transferred among bacterial cells, and (2) the presence of multiple drug resistance determinants.

8 Lethality in all microenvironments eliminates the allele in one generation; lethality in half the microenvironments decreases the allele frequency by $q' = q/(2 - q)$.

9 In Batesian mimicry, a palatable mimic resembles an unpalatable model, and there is frequency-dependent selection against the mimic becoming too common. In Müllerian mimicry, unpalatable species mimic each other,

and there is frequency-dependent selection because a high frequency of mimicry is mutually beneficial.

10 In a haploid with constant selection, $(q_n/p_n) = (q_0/p_0)(1 - s)^n = (q_0/p_0)(1 - 0.01)^{10} = (q_0/p_0)(0.9044)$. With variable selection, $(q_{10}/p_{10}) = (q_0/p_0)(0.9000)$, so constant selection is the more effective. This is because the effects get compounded every generation.

11 The sister must receive the same allele from the father, and the other allele will be identical by descent with probability 1/2.

12 The early reproducing population because (1) the early reproduction is compounded through time, giving a somewhat faster rate of growth, and (2) delayed reproduction risks prereproductive mortality.

13 The summation theorem implies that all enzymes control the flux to a greater or lesser extent, and large control coefficients for some steps must imply small control coefficients for others.

14 For $a = 10$, $(1.1)(10)/[1 + (1.1)(10)] = 0.9167$ versus $10/11 = 0.9091$, ratio = 1.008, so fitness increase = 0.8%. For $a = 100$, fitness increase = 0.09%.

15 F_{XY} is the probability that a random gamete from X contains an allele that is identical by descent with one from Y; $(1 + F_X)/2$ is the probability that two gametes from X contain alleles that are identical by descent. The coefficient $r = 2F_{XY}/(1 + F_X)$ is the ratio.

16 Fitnesses of (A, a) are (1, 1) in a proportion $1 - s$ of environments and (1, 0) in a proportion s. Therefore, $\Delta q_{sel} = (1 - s)(0) + s(-q) = -qs$. Rate of origin of new a alleles by mutation $\Delta q_{mut} = p\mu$. Setting $\Delta q_{sel} = \Delta q_{mut}$ gives $-qs = p\mu$ or $q = \mu/(s + \mu) \approx \mu/s$.

17 Allele frequency alternates between $(-1 + \sqrt{5})/2$ and $(3 - \sqrt{5})/2$. Let q = frequency of lethal allele in the first generation, then $q' = q/(1 + q)$. In the next generation the other allele is lethal, so $q'' = 1/(2 - q') = (1 + q)/(2 + q)$. Set $q'' = q$ and solve for $q = (-1 + \sqrt{5})/2$, $p = (3 - \sqrt{5})/2$. Note $q' = (-1 + \sqrt{5})/(1 + \sqrt{5}) = (3 - \sqrt{5})/2$, so the allele frequency alternates.

18 Frequencies after one generation are $y_1' = (y_1/2)/T$ and $y_2' = (ky_2)/T$, where $T = (y_1/2) + y_2$. Therefore, $y_2'/y_1' = (y_2/y_1)(2k)$, or $y_2(t)/y_1(t) = [y_2(0)/y_1(0)](2k)^t$. The ultimate fate is fixation of the meiotic drive Y. With fitness of the meiotic drive male of w, then $y_2'/y_1' = (y_2/y_1)(2kw)$, and frequencies do not change if $w = 1/(2k)$. This does not result in a stable equilibrium, however. Any small change in w or k will lead to fixation or loss of the driven Y chromosome.

19 In this situation, $r^*N[1 - (N/K^*)] - rN[1 - (N/K)]$ is equivalent to $w_i - \bar{w}$.

20 Depends on γ because the chance of A mating with the wrong species increases as the frequency of A decreases. If $\gamma = 1/2$, $\Delta p = pqr/2$, which is approximately the same change as in a diploid additive viability model with fitnesses $1 + r$, $1 + r/2$, and 1 for the three genotypes.

Bibliography

Abramowitz, M. and I.A. Stegun. 1972. *Handbook of Mathematical Functions, with Formulas, Graphs, and Mathematical Tables*. Dover Publications, New York.

Allard, R.W., A.L. Kahler and B.S. Weir. 1972. The effect of selection on esterase allozymes in a barley population. Genetics 72:489–503.

Allison, A.C. 1964. Polymorphism and natural selection in human populations. Cold Spring Harbor Symp. Quant. Biol. 29:137–149.

Altenberg, L. and M.W. Feldman. 1987. Selection, generalized transmission and the evolution of modifier genes. I. The reduction principle. Genetics 117:559–572.

Anderson, E.S. 1969. Ecology and epidemiology of transferable drug resistance. pp. 102–115. *In* G.E.W. Wolstenholme and M. O'Connor (eds.), *Bacterial Episomes and Plasmids*. Little, Brown and Co., Boston.

Anderson, W.W. 1969. Polymorphism resulting from the mating advantage of rare male genotypes. Proc. Natl. Acad. Sci. USA 64:190–197.

Anderson, W.W. 1971. Genetic equilibrium and population growth under density-regulated selection. Am. Nat. 105:489–498.

Anderson, W.W., J. Arnold, S.A. Sammons and D.G. Yardley. 1986. Frequency-dependent viabilities of *Drosophila pseudoobscura* karyotypes. Heredity 56:7–17.

Anderson, W.W. and C.J. Brown. 1984. A test for rare male mating advantage with *Drosophila pseudoobscura* karyotypes. Genetics 107:577–589.

Anderson, W.W. and C.E. King. 1970. Age-specific selection. Proc. Natl. Acad. Sci. USA 66:780–786.

Anderson, W.W., C. Oshima, T. Watanabe, Th. Dobzhansky and O. Pavlovsky. 1968. Genetics of natural populations. XXXIX. A test of the possible influence of two insecticides on the chromosomal polymorphism in *Drosophila pseudoobscura*. Genetics 58:423–434.

Antonovics, J. 1968. Evolution in closely adjacent plant populations. V. Evolution of self-sterility. Heredity 23:219–238.

Antonovics, J. and A.D. Bradshaw. 1970. Evolution in closely adjacent plant populations. VII. Clinal patterns at a mine boundary. Heredity 25:349–362.

Antonovics, J., A.D. Bradshaw and R.G. Turner. 1971. Heavy metal tolerance in plants. Adv. Ecol. Res. 7:1–85.

Anxolabéhère, D. 1980. The influence of sexual and larval selection on the maintenance of polymorphism at the *sepia* locus in *Drosophila melanogaster*. Genetics 95:743–755.

Anxolabéhère, D., J.M. Goux and G. Periquet. 1982. A bias in estimation of viabilities from competition experiments. Heredity 48:271–282.

Anxolabéhère, D., M.G. Kidwell and G. Periquet. 1988. Molecular characteristics of diverse populations are consistent with the hypothesis of a recent invasion of *Drosophila melanogaster* by mobile *P* elements. Mol. Biol. Evol. 5:252–269.

Aquadro, C.F., S.F. Desse, M.M. Bland, C.H. Langley and C.C. Laurie-Ahlberg. 1986. Molecular population genetics of the alcohol dehydrogenase gene region of *Drosophila melanogaster*. Genetics 114:1165–1190.

Arnheim, N, M. Krystal, M.R. Schmickel, G. Wilson, O. Ryder and E. Zimmer. 1980. Molecular evidence for genetic exchanges among ribosomal genes on nonhomologous chromosomes in man and apes. Proc. Natl. Acad. Sci. USA 77:7323–7327.

Arnold, S.J. and M.J. Wade. 1984a. On the measurement of natural and sexual selection: Theory. Evolution 38:709–719.

Arnold, S.J. and M.J. Wade. 1984b. On the measurement of natural and sexual selection: Applications. Evolution 38:720–734.

Asmussen, M.A., J. Arnold and J.C. Avise. 1987. Definition and properties of disequilibrium statistics for associations between nuclear and cytoplasmic genotypes. Genetics 115:755–768.

Atchley, W.R. 1984. Ontogeny, timing of development, and genetic variance-covariance structure. Am. Nat. 123:519–540.

Atchley, W.R., J.J. Rutledge and D.E. Cowley. 1982. A multivariate statistical analysis of direct and correlated response to selection in the rat. Evolution 36:677–698.

Avise, J.C. 1986. Mitochondrial DNA and the evolutionary genetics of higher animals. Phil. Trans. R. Soc. Lond. B 312:325–342.

Avise, J.C., E. Bermingham, L.G. Kessler and N.C. Saunders. 1984. Characterization of mitochondrial DNA variability in a hybrid swarm between subspecies of bluegill sunfish (*Lepomis macrochirus*). Evolution 38:931–941.

Avise, J.C., C. Giblin-Davidson, J. Laerm, J.C. Patton and R.A. Lansman. 1979a. Mitochondrial DNA clones and matriarchal phylogeny within and among geographic populations of the pocket gopher, *Geomys oinetis*. Proc. Natl. Acad. Sci. USA 76:6694–6698.

Avise, J.C., M.H. Smith and R.K. Selander. 1979b. Biochemical polymorphism and systematics in the genus *Peromyscus*. VII. Geographic differentiation in members of the *truei* and *maniculatus* species groups. J. Mamm. 60:177–192.

Avise, J.C., J.F. Shapira, S.W. Daniel, C.F. Aquadro and R.A. Lansman. 1983. Mitochondrial DNA differentiation during the speciation process in *Peromyscus*. Mol. Biol. Evol. 1:38–56.

Axelrod, R. and W.D. Hamilton. 1981. The evolution of cooperation. Science 211:1390–1396.

Ayala, F.J. and C.A. Campbell. 1974. Frequency-dependent selection. Annu. Rev. Ecol. Syst. 5:115–138.

Ayala, F.J. and M.L Tracey. 1974. Genetic differentiation within and between species of the *Drosophila willistoni* group. Proc. Natl. Acad. Sci. USA 71:999–1003.

Baird, M., I. Ballazs, A. Giusti, L. Miyazaki, L. Nicholas, K. Wexler, E. Kanter, J. Glassberg, F. Allen, P. Rubenstein and L. Sussman. 1986. Allele frequency distribution of two highly polymorphic DNA sequences in three ethnic groups and its applicability to the determination of paternity. Am. J. Hum. Gen. 39:489–501.

Baker, R.L., A.B. Chapman and R.T. Wardell. 1975. Direct response to selection for postweaning weight gain in the rat. Genetics 80:171–189.

Baker, W.K. 1975. Linkage disequilibrium over space and time in natural populations of *Drosophila montana*. Proc. Natl. Acad. Sci. USA 72:4095–4099.

Barker, J.S.F. 1979. Interlocus interactions: A review of experimental evidence. Theor. Pop. Biol. 16:323–346.

Barker, J.S.F. and P.D. East. 1980. Evidence for selection following perturbation of allozyme frequencies in a natural population of *Drosophila*. Nature 284:166–168.

Barker, J.S.F., P.D. East and B.S. Weir. 1986. Temporal and microgeographic variation in allozyme frequencies in a natural population of *Drosophila buzzatii*. Genetics 112:577–611.

Barnes, P.T. and C.C. Laurie-Ahlberg. 1986. Genetic variability of flight metabolism in *Drosophila melanogaster*. III. Effects of *Gpdh* allozymes and environmental temperature on power output. Genetics 112:267–294.

Barton, N.H. and M. Slatkin. 1986. A quasi-equilibrium theory of the distribution of rare alleles in a subdivided population. Heredity 56:409–416.

Barton, N.H. and M. Turelli. 1987. Adaptive landscapes, genetic distance and the evolution of quantitative characters. Genet. Res. 49:157–173.

Beckmann, J.S. and M. Soller. 1983. Restriction fragment length polymorphism in genetic improvement: Methodologies, mapping and costs. Theor. Appl. Genet. 67:35–43.

Beckmann, J.S. and M. Soller. 1986. Restriction fragment length polymorphisms and genetic improvement of agricultural species. Euphytica 35:111–124.

Bell, G. 1982. *The Masterpiece of Nature: The Evolution and Genetics of Sexuality*. Croom Helm, London.

Benjamin, D.C., J.A. Berzofsky, I.J. East, F.R.N. Gurd, C. Hannum, S.J. Leach, E. Margoliash, J.G. Michael, A. Miller, E.M. Prager, M. Reichlin, E.E. Sercarz, S.J. Smith-Gill, P.E. Todd and A.C. Wilson. 1984. The antigenic structure of proteins: A reappraisal. Annu. Rev. Immunol. 2:67–101.

Bennett, J.H. 1957. Selectively balanced polymorphism at a sex-linked locus. Nature 180:1363–1364.

Bennett, J.H. and F.E. Binet. 1956. Association between Mendelian factors with mixed selfing and random mating. Heredity 10:51–56.

Bennett, J.H. and C.R. Oertel. 1965. The approach to a random association of genotypes with random mating. J. Theor. Biol. 9:67–76.

Benveniste, R.E. 1985. The contributions of retroviruses to the study of mammalian evolution. pp. 359–417. *In* R.J. MacIntyre (ed.), *Molecular Evolutionary Genetics*. Plenum, New York.

Benveniste, R. and J. Davies. 1973. Aminoglycoside antibiotic inactivating enzymes in *Actinomycetes* similar to those present in clinical isolates of antibiotic-resistant bacteria. Proc. Natl. Acad. Sci. USA 70:2276–2280.

Berg, D.E., J. Davies, B. Allet and J.-D. Rochaix. 1975. Transposition of R-factor genes to bacteriophage lambda. Proc. Natl. Acad. Sci. USA 72:3628–3632.

Bermingham, E., T. Lamb and J.C. Avise. 1986. Size polymorphism and heteroplasmy in the mitochondrial DNA of lower vertebrates. J. Hered. 77:249–252.

Bernstein, F. 1930. Über die Erblichkeit der Blutgrüppen. Z. Absat. u. Vererb. 54:400–426.

Bernstein, S.C., L.H. Throckmorton and J.L. Hubby. 1973. Still more genetic variability in natural populations. Proc. Natl. Acad. Sci. USA 70:3928–3931.

Beverly, S.M. and A.C. Wilson. 1985. Ancient origin for Hawaiian Drosophilidae inferred from protein comparisons. Proc. Natl. Acad. Sci. USA 82:4753–4757.

Biémont, C., A. Aouar and C. Arnault. 1987. Genome reshuffling of the *copia* element in an inbred line of *Drosophila melanogaster*. Nature 329:742–744.

Bijlsma, R. 1978. Polymorphism at the *G6PD* and *6PGD* loci in *Drosophila melanogaster*. II. Evidence for interaction in fitness. Genet. Res. 31:227–237.

Bijlsma, R. and W. van Delden. 1977. Polymorphism at the *G6PD* and *6PGD* loci in *Drosophila melanogaster*. I. Evidence for selection in experimental populations. Genet. Res. 30:221–236.

Bingham, P.M., M.G. Kidwell and G.M. Rubin. 1982. The molecular basis of *P–M* hybrid dysgenesis: The role of the *P* element, a *P*-strain-specific transposon family. Cell 29:995–1004.

Birky, C.W., T. Maruyama and P. Fuerst. 1983. An approach to population and evolutionary genetic theory for genes in mitochondria and chloroplasts, and some results. Genetics 103:513–527.

Birky, C.W. and J.B. Walsh. 1988. Effects of linkage on rates of molecular evolution. Proc. Natl. Acad. Sci. USA 85:6414–6418.

Bishop, J.A. and L.M. Cook. 1975. Moths, melanism and clean air. Sci. Am. 232:90–99.

Bishop, J.M. 1981. Enemies within: The genesis of retrovirus oncogenes. Cell 23:5–6.

Bishop, J.M. 1982. Oncogenes. Sci. Am. 246:80–92.

Bodmer, W.F. 1965. Differential fertility in population genetics models. Genetics 51:411–424.

Bodmer, W.F. and J. Felsenstein. 1967. Linkage and selection: Theoretical analysis of the deterministic two-locus random mating model. Genetics 57:237–265.

Bodmer, M. and M. Ashburner. 1984. Conservation and change in the DNA sequences coding for alcohol dehydrogenase in sibling species of *Drosophila*. Nature 309:425–430.

Boerwinkle, S. Visvikis, D. Welsh, J. Steinmetz, S.M. Hanash and C.F. Sing. 1987. The use of measured genotype information in the analysis of quantitative phenotypes in man. II. The role of the apolipoprotein E polymorphism in determining levels, variability and covariability of cholesterol, betalipoprotein and triglycerides in a samples of unrelated individuals. Am. J. Med. Genet. 27:567–582.

Boggs, C.L.. 1981. Selection pressures affecting male nutrient investment at mating in heliconiine butterflies. Evolution 35:931–940.

Bohren, B.B., W.G. Hill and A. Robertson. 1966. Some observations on asymmetrical correlated responses to selection. Genet. Res. 7:44–57.

Boller, E.F. and G.L. Bush. 1973. Evidence for genetic variation in populations of the European cherry fruit fly *Rhagoletis cerasi* (Diptera: Tephritidae) based on physiological parameters and hybridization experiments. Entomol. Exp. Appl. 17:279–293.

Bonhomme, F. and R.K. Selander. 1978. Estimating total genic diversity in the house mouse. Biochem. Genet. 16:287–298.

Bonnell, M.L. and R.K. Selander. 1974. Elephant seals: Genetic variation and near extinction. Science 184:908–909.

Boorman, S.A. and P.R. Levitt. 1980. *The Genetics of Altruism*. Academic Press, New York.

Botstein, D. R.L. White, M. Skolnick and R.W. Davis. 1980. Construction of a genetic linkage map in man using restriction fragment length polymorphisms. Am. J. Hum. Gen. 32:314–331.

Boursot, P., H. Yonekawa and F. Bonhomme. 1987. Heteroplasmy in mice with deletion of a large coding region of mitochondrial DNA. Mol. Biol. Evol. 4:46–55.

Boussy, I.A. and M.G. Kidwell. 1987. The *P–M* hybrid dysgenesis cline in eastern Australian *Drosophila melanogaster*: Discrete *P*, *Q*, and *M* regions are nearly contiguous. Genetics 115:737–745.

Boyd, R. and J.P. Lorberbaum. 1987. No pure strategy is evolutionarily stable in the repeated Prisoner's Dilemma game. Nature 327:58–59.

Bregliano, J.-C. and M.G. Kidwell. 1983. Hybrid dysgenesis determinants. pp. 363–410. *In* J.A. Shapiro (ed.), *Mobile Genetic Elements*, Academic Press, New York.

Briggs, G.E. and J.B.S. Haldane. 1925. A note on the kinetics of enzyme action. Biochem. J. 19:338–339.

Britten, R.J. 1986. Rates of DNA sequence evolution differ between taxonomic groups. Science 231:1393–1398.

Brittnacher, J.G. and B. Ganetzky. 1984. On the components of segregation distortion in *Drosophila melanogaster*. III. Nature of enhancer of *SD*. Genetics 107:423–434.

Brookfield, J.F.Y. 1986. A model for DNA sequence evolution within transposable element families. Genetics 112:393–407.

Brower, J.V.Z. 1958. Experimental studies of mimicry in some North American butterflies. Part I. The Monarch, *Danaus plexippus* and Viceroy, *Limenitis archippus archippus*. Evolution 12:32–47.

Brower, L.P., L.M. Cook and H.J. Croze. 1967. Predator response to artificial Batesian mimics released in a neotropical environment. Evolution 21:11–23.

Brown, A.H.D. 1975. Sample sizes required to detect linkage disequilibrium between two or three loci. Theor. Pop. Biol. 8:184–201.

Brown, A.H.D. and M.T. Clegg. 1983. Analysis of variation in related DNA sequences. pp. 107–132. *In* B. Weir (ed.), *Statistical Analysis of DNA Sequence Data*. Marcel Dekker, New York.

Brown, A.H.D. and M.W. Feldman. 1981. Population structure of multilocus associations. Proc. Natl. Acad. Sci. USA 78:5913–5916.

Brown, A.H.D., M.W. Feldman and E. Nevo. 1980. Multilocus structure of natural populations of *Hordeum spontaneum*. Genetics 96:523–536.

Brown, W.M. 1980. Polymorphism in mitochondrial DNA of humans as revealed by restriction endonuclease analysis. Proc. Natl. Acad. Sci. USA 77:3605–3609.

Brown, W.M. 1983. Evolution of animal mitochondrial DNA. pp. 62–88. *In* M. Nei and R.K. Koehn (eds.), *Evolution of Genes and Proteins*, Sinauer Associates, Sunderland, MA.

Brown, W.M. 1985. The mitochondrial genome of animals. pp. 95–130. *In* R.J. MacIntyre (ed.), *Molecular Evolutionary Genetics*. Plenum, New York.

Brown, W.M., M. George Jr. and A.C. Wilson. 1979. Rapid evolution of animal mitochondrial DNA. Proc. Natl. Acad. Sci. USA 76:1967–1971.

Brownell, E., M. Krystal and N. Arnheim. 1983. Structure and evolution of human and African ape rDNA pseudogenes. Mol. Biol. Evol. 1:29–37.

Bruck, D. 1957. Male segregation ratio advantage as a factor in maintaining lethal alleles in wild populations of house mice. Proc. Natl. Acad. Sci. USA 43:152–158.

Bryan, G.J., J.W. Jacobson and D.L Hartl. 1987. Heritable somatic excision of a *Drosophila* transposon. Science 235:1636–1638.

Bryant, E.H., A. Kence and K.T. Kimball. 1980. A rare-male advantage in the housefly induced by wing clipping and some general considerations for *Drosophila*. Genetics 96:975–993.

Bucheton, A. R. Paro, H.M. Sang, A. Pelisson and D.J. Finnegan. 1984. The molecular basis of *I–R* hybrid dysgenesis in *Drosophila melanogaster*: Identification, cloning and properties of the *I* factor. Cell 38:153–163.

Bull, J.J. 1983. *Evolution of Sex Determining Mechanisms*. Benjamin/Cummings, Menlo Park, CA.

Bulmer, M.G. 1970. *The Biology of Twinning in Man*. Oxford University Press, Oxford.

Bulmer, M.G. 1980. *The Mathematical Theory of Quantitative Genetics*. Oxford University Press, Oxford.

Bumpus, H.C. 1899. The elimination of the unfit as illustrated by the introduced sparrow *Passer domesticus*. Woods Hole Marine Biological Station, Biological Lectures 6:209–226.

Bundgaard, J. and F.B. Christiansen. 1972. Dynamics of polymorphisms. I. Selection components in an experimental population of *Drosophila melanogaster*. Genetics 71:439–460.

Buonagurio, D.A., S. Nakada, J.D. Parvin, M. Krystal, P. Palese and W.M. Fitch. 1986. Evolution of human influenza A viruses over 50 years: Rapid, uniform rate of change in *NS* gene. Science 232:980–982.

Buri, P. 1956. Gene frequency in small populations of mutant *Drosophila*. Evolution 10:367–402.

Burr,B., F.A. Burr, K.H. Thompson, M.C. Albertson and C.W. Stuber. 1988. Gene mapping with recombinant inbreds in maize. Genetics 118:519–526.

Burton, R.S. 1987. Differentiation and integration of the genome in populations of the marine copepod *Tigriopus californicus*. Evolution 41:504–513.

Burton, R.S. and M.W. Feldman. 1981. Population genetics of *Tigriopus californicus*. II. Differentiation among neighboring populations. Evolution 35:1192–1205.

Burton, R.S. and M.W. Feldman. 1983. Physiological effects of an allozyme polymorphism: Glutamate-pyruvate transaminase and response to hyperosmotic stress in the copepod *Tigriopus californicus*. Biochem. Genet. 21:239–251.

Burton, R.S. and A.R. Place. 1986. Evolution of selective neutrality: Further considerations. Genetics 114:1033–1036.

Bush, G.L. 1975. Modes of animal speciation. Annu. Rev. Ecol. Syst. 6:339–364.

Bush, G.L., S.M. Case, A.C. Wilson and J.L. Patton. 1977. Rapid speciation and chromosomal evolution in mammals. Proc. Natl. Acad. Sci. USA 74:3942–3946.

Bush, R.M., P.E. Smouse and F.T. Ledig. 1987. The fitness consequences of multiple-locus heterozygosity: The relationship between heterozygosity and growth rate in pitch pine (*Pinus rigida*). Evolution 41:787–798.

Caisse, M. and J. Antonovics. 1978. Evolution in closely adjacent plant populations. IX. Evolution of reproductive isolation in clinal populations. Heredity 40:371–384.

Campbell, A.M. 1969. *Episomes*. Harper & Row, New York.

Cann, R.L. and A.C. Wilson. 1983. Length mutations in human mitochondrial DNA. Genetics 104:699–711.

Cann, R.L., M. Stoneking and A.C. Wilson. 1987. Mitochondrial DNA and human evolution. Nature 325:31–36.

Cannings, C. 1967. Equilibrium, convergence and stability at a sex-linked locus under natural selection. Genetics 56:613–618.

Carson, H. 1968. The population flush and its genetic consequences. pp. 123–137. *In* R.C. Lewontin (ed.), *Population Biology and Evolution*. Syracuse University Press, Syracuse, NY.

Carson, H.L. 1971. Speciation and the founder principle. Stadler Genetic Symposium 3:51–70.

Carson, H.L. 1975. The genetics of speciation at the diploid level. Am. Nat. 109:83–92.

Carson, H.L. 1978. Speciation and sexual selection in Hawaiian *Drosophila*. pp. 93–107. *In* P.F. Brussard (ed.), *Ecological Genetics: The Interface*. Springer-Verlag, New York.

Carson, H.L. and P.J. Bryant. 1979. Change in a secondary sexual character as evidence of incipient speciation in *Drosophila sylvestris*. Proc. Natl. Acad. Sci. USA 76:1929–1932.

Carson, H.L., D.E. Hardy, H.T. Spieth and W.S. Stone. 1970. The evolutionary biology of the Hawaiian Drosophilidae. Evol. Biol. 4 (Suppl):437–544.

Carson, H.L. and K.Y. Kaneshiro. 1976. *Drosophila* of Hawaii: Systematics and ecological genetics. Annu. Rev. Ecol. Syst. 7:311–346.

Carson, H.L. and A.R. Templeton. 1984. Genetic revolutions in relation to speciation phenomena: The founding of new populations. Annu. Rev. Ecol. Syst. 15:97–131.

Carter, L.O. 1961. The inheritance of congenital pyloric stenosis. Brit. Med. Bull. 17:251–254.

Castle, W.E. 1921. An improved method of estimating the number of genetic factors concerned in cases of blending inheritance. Science 54:223.

Cavalli-Sforza, L.L. 1966. Population structure and human evolution. Proc. Roy. Soc. Lond. B 164:362–379.

Cavalli-Sforza, L.L. and W.F. Bodmer. 1971. *The Genetics of Human Populations*. W.H. Freeman, San Francisco.

Cavalli-Sforza, L.L. and M.W. Feldman. 1978. Darwinian selection and "altruism." Theor. Pop. Biol. 14:268–280.

Cavender, D.E., D.K. Wagener, B.S. Rabin et al. 1984. The Pittsburgh insulin-dependent diabetes mellitus (IDDM) study: HLA antigens and haplotypes as risk factors for the development of IDDM in IDDM patients and their siblings. J. Chronic Dis. 37:555–568.

Cavener, D.R. and M.T. Clegg. 1981a. Multigenic response to ethanol in *Drosophila melanogaster*. Evolution 35:1–10.

Cavener, D.R. and M.T. Clegg. 1981b. Evidence for biochemical and physiological differences between enzyme genotypes in *Drosophila melanogaster*. Proc. Natl. Acad. Sci. USA 78:4444–4447.

Chakraborty, R., P.A. Fuerst and M. Nei. 1980. Statistical studies on protein polymorphism in natural populations. III. Distribution of allele frequencies and the number of alleles per locus. Genetics 94:1039–1063.

Chakraborty, R. and M. Nei. 1977. Bottleneck effects on average heterozygosity and genetic distance with the stepwise mutation model. Evolution 31:347–356.

Champion, A.B., E.M. Prager, D. Wachter and A.C. Wilson. 1974. Microcomplement fixation. pp. 397–416. *In* C.A. Wright (ed.), *Biochemical and Immunological Taxonomy of Animals*. Academic Press, London.

Chao, L., C. Vargas, B.B. Spear and E.C. Cox. 1983. Transposable elements as mutator genes in evolution. Nature 303:633–635.

Chapman, R.W., J.C. Stephens, R.A. Lansman and J.C. Avise. 1982. Models of mitochondrial DNA transmission genetics and evolution in higher eukaryotes. Genet. Res. 40:41–57.

Chappell, M.A., J.P. Hayes and L.R.G. Snyder. 1988. Hemoglobin polymorphisms in deer mice (*Peromyscus maniculatus*): Physiology of β-globin variants and α-globin recombinants. Evolution 42:681–688.

Charlesworth, B. 1970. Selection in populations with overlapping generations. I. The use of Malthusian parameters in population genetics. Theor. Pop. Biol. 1:352–370.

Charlesworth, B. 1971. Selection in density-regulated populations. Ecology 52:469–474.

Charlesworth, B. 1972. Selection in populations with overlapping generations. III. Conditions for genetic equilibrium. Theor. Pop. Biol. 3:377–395.

Charlesworth, B. 1973. Selection in populations with overlapping generations. V. Natural selection and life histories. Am. Nat. 107:303–311.

Charlesworth, B. 1974. Selection in populations with overlapping generations. VI. Rates of change of gene frequency and population growth rate. Theor. Pop. Biol. 6:108–133.

Charlesworth, B. 1978. Some models of the evolution of altruistic behavior between siblings. J. Theor. Biol. 72:297–319.

Charlesworth, B. 1980. *Evolution in Age-Structured Populations*. Cambridge University Press, Cambridge.

Charlesworth, B. 1985a. Recombination, genome size and chromosome number. pp. 489–513. *In* T. Cavalier-Smith (ed.), *The Evolution of Genome Size*. Wiley, New York.

Charlesworth, B. 1985b. The population genetics of transposable elements. pp. 213–222. *In* T. Ohta and K.-I. Aoki (eds.), *Population Genetics and Molecular Evolution*. Springer-Verlag, Berlin.

Charlesworth, B. and D. Charlesworth. 1983. The population dynamics of transposable elements. Genet. Res. 42:1–27.

Charlesworth, B., C. Langley and W. Stephan. 1986. The evolution of restricted recombination and the accumulation of repeated DNA sequences. Genetics 112:947–962.

Charlesworth, D. and B. Charlesworth. 1975. Theoretical genetics of Batesian mimicry. I. Single-locus models. J. Theor. Biol. 55:283–303.

Charlesworth, D. and B. Charlesworth. 1976a. Theoretical genetics of Batesian mimicry. II. Evolution of supergenes. J. Theor. Biol. 55:305–324.

Charlesworth, D. and B. Charlesworth. 1976b. Theoretical genetics of Batesian mimicry. III. Evolution of dominance. J. Theor. Biol. 55:325–337.

Charnov, E.L. 1982. *The Theory of Sex Allocation*. Princeton University Press, Princeton, NJ.

Charnov, E.L. and J. Bull. 1977. When is sex environmentally determined? Nature 266:828–830.

Christiansen, F.B. 1974. Sufficient conditions for protected polymorphism in a subdivided population. Am. Nat. 108:157–166.

Christiansen, F.B. 1975. Hard and soft selection in a subdivided population. Am. Nat. 109:11–16.

Christiansen, F.B., J. Bundgaard and J.S.F. Barker. 1977. On the structure of fitness estimates under post-observational selection. Evolution 31:843–853.

Christiansen, F.B. and O. Frydenberg. 1973. Selection component analysis of natural polymorphisms using population samples including mother–offspring combinations. Theor. Pop. Biol. 4:425–445.

Christiansen, F.B. and O. Frydenberg. 1974. Geographical patterns of four polymorphisms in *Zoarces viviparus* as evidence of selection. Genetics 77:765–770.

Christiansen, F.B., O. Frydenberg and V. Simonsen. 1973. Genetics of *Zoarces* populations. IV. Selection component analysis of an esterase polymorphism using population samples including mother–offspring combinations. Hereditas 73: 291–304.

Christiansen, F.B., O. Frydenberg and V. Simonsen. 1977. Genetics of *Zoarces* populations. X. Selection component analysis of the *Est*III polymorphism using samples of successive cohorts. Hereditas 87:129–150.

Clark, A.G. 1984. Natural selection with nuclear and cytoplasmic transmission. I. A deterministic model. Genetics 107:679–701.

Clark, A.G. 1986. Partitioning genetic variation in viability among lines of *Drosophila melanogaster*. Am. Nat. 128:263–271.

Clark, A.G. 1987a. A test of multilocus interaction in *Drosophila melanogaster*. Am. Nat. 130:283–299.

Clark, A.G., 1987b. Neutrality tests of highly polymorphic restriction fragment length polymorphisms. Am. J. Hum. Gen. 41:948–956.

Clark, A.G. 1988. Deterministic theory of heteroplasmy. Evolution 42:621–625.

Clark, A.G. and J. Bundgaard. 1984. Selection components in background replacement lines of *Drosophila*. Genetics 108:181–200.

Clark, A.G. and W.W. Doane. 1984. Interactions between the *Amylase* and *adipose* chromosomal regions of *Drosophila melanogaster*. Evolution 38:957–982.

Clark, A.G. and M.W. Feldman. 1981a. The estimation of epistasis in components of fitness in experimental populations of *Drosophila melanogaster*. II. Assessment of meiotic drive, viability, fecundity and sexual selection. Heredity 46:347–377.

Clark, A.G. and M.W. Feldman. 1981b. Density-dependent fertility selection in experimental populations of *Drosophila melanogaster*. Genetics 98:849–869.

Clark, A.G. and M.W. Feldman. 1986. A numerical simulation of the one-locus multiple-allele fertility model. Genetics 113:161–176.

Clark, A.G., M.W. Feldman and F.B. Christiansen. 1981. The estimation of epistasis in components of fitness in experimental populations of *Drosophila melanogaster*. I. A two-stage maximum likelihood model. Heredity 46:321–346.

Clark, A.G. and L.E. Keith. 1988. Variation among extracted lines of *Drosophila melanogaster* in triacylglycerol and carbohydrate storage. Genetics 119:595–607.

Clark, A.G. and E.M.S. Lyckegaard. 1988. Natural selection with nuclear and cytoplasmic transmission. III. Joint analysis of segregation and mtDNA in *Drosophila melanogaster*. Genetics 118:471–481.

Clarke, B. and J. Murray. 1969. Ecological genetics and speciation in land snails of the genus *Partula*. Biol. J. Linn Soc. 1:31–42.

Clarke, G.M. and J.A. McKenzie. 1987. Developmental stability of insecticide resistant phenotypes in blowfly: A result of canalizing natural selection. Nature 325:345–346.

Clayton, G.A., J.A. Morris and A. Robertson. 1953. Selection for abdominal chaetae in a large population of *Drosophila melanogaster*. pp. 7–15. *In Symposium on the Genetics of Population Structure*. Premiata Tipografia Successori Filli Fusi, Pavia, Italy.

Clayton, G.A. and A. Robertson. 1955. Mutation and quantitative variation. Am. Nat. 89:151–158.

Clayton, G. and A. Robertson. 1957. An experimental check on quantitative genetical theory. II. Long-term effects of selection. J. Genet. 55:152–170.

Clegg, M.T. and R.W. Allard. 1972. Patterns of genetic differentiation in the slender wild oat species *Avena barbata*. Proc. Natl. Acad. Sci. USA 69:1820–1824.

Clegg, M.T., R.W. Allard and A.L. Kahler. 1972. Is the gene the unit of selection? Evidence from two experimental plant populations. Proc. Natl. Acad. Sci. USA 69:2474–2478.

Clegg, M.T., A.L Kahler and R.W. Allard. 1978. Estimation of life cycle components of selection in an experimental plant population. Genetics 89:765–792.

Clegg, M.T., J.F. Kidwell and C.R. Horch. 1980. Dynamics of correlated genetic systems. V. Rates of decay of linkage disequilibria in experimental populations of *Drosophila melanogaster*. Genetics 94:217–234.

Clegg, M.T., J.F. Kidwell and M.G. Kidwell. 1978. Dynamics of correlated genetic systems. III. Behavior of chromosomal segments under lethal selection. Genetica 88:95–106.

Clegg, M.T., J.F. Kidwell, M.G. Kidwell and N.J. Daniel. 1976. Dynamics of correlated genetic systems. I. Selection in the region of the glued locus of *Drosophila melanogaster*. Genetics 83:793–810.

Clowes, R.C. 1972. Molecular structure of bacterial plasmids. Bacteriol. Rev. 36:361–405.

Cluster, P.D., D. Marinkovic, R.W. Allard and F.J. Ayala. 1987. Correlations between developmental rates, enzyme activities, ribosomal DNA spacer-length phenotypes, and adaptation in *Drosophila melanogaster*. Proc. Natl. Acad. Sci. USA 84:610–614.

Cockerham, C.C. 1969. Variance of gene frequencies. Evolution 23:72–84.

Cockerham, C.C. 1973. Analysis of gene frequencies. Genetics 74:679–700.

Cockerham, C.C. 1986. Modifications in estimating the number of genes for a quantitative character. Genetics 114:659–664.

Cockerham, C.C., P.M. Burrows, S.S. Young and T. Prout. 1972. Frequency-dependent selection in randomly mating populations. Am. Nat. 106:493–515.

Cockerham, C.C. and J.O. Rawlings. 1967. Apparent heterosis of a neutral gene with inbreeding. Ciencia e Cultura 19:89–94.

Cockerham, C.C. and H. Tachida. 1987. Evolution and maintenance of quantitative genetic variation by mutations. Proc. Natl. Acad. Sci. USA 84:6205–6209.

Cockerham, C.C. and B.S. Weir. 1968. Sib mating with two linked loci. Genetics 60:629–640.

Cockerham, C.C. and B.S. Weir. 1977. Digenic descent measures for finite populations. Genet. Res. 30:121–147.

Coen, E.S., J.M. Thoday and G.A. Dover. 1982. Rate of turnover of the structural variants in the rDNA gene of *Drosophila melanogaster*. Nature 295:564–568.

Cohan, F.M. and J.-D. Graf. 1985. Latitudinal cline in *Drosophila melanogaster* knockdown resistance to ethanol fumes and for rates of response to selection for further resistance. Evolution 39:278–293.

Cohn, V.H., M.A. Thompson and G.P. Moore. 1984. Nucleotide sequence comparison of the *Adh* gene in three drosophilids. J. Mol. Evol. 20:31–37.

Cook, L.M. 1965. Inheritance of shell size in the snail *Arianta arbustorum*. Evolution 19:86–94.

Cook, L.M., G.S. Mani and M.E. Varley. 1986. Post-industrial melanism in the peppered moth. Science 231:622–623.

Cook, S.C.A., C. Lefebvre and T. McNeilly. 1972. Competition between metal tolerant and normal plant populations on normal soil. Evolution 26:366–372.

Cotterman, C.W. 1940. A calculus for statistico-genetics. Ph.D. thesis, Ohio State University, Columbus.

Coyne, J.A. 1983. Genetic basis of differences in genital morphology among three sibling species of *Drosophila*. Evolution 37:1101–1118.

Coyne, J.A. 1984. Genetic basis of male sterility in hybrids between two closely related species of *Drosophila*. Proc. Natl. Acad. Sci. USA 81:4444–4447.

Coyne, J.A. 1985. The genetic basis of Haldane's rule. Nature 314:736–738.

Coyne, J.A. and E. Beecham. 1987. Heritability of two morphological characters within and among natural populations of *Drosophila melanogaster*. Genetics 117:727–737.

Coyne, J.A., A.A. Felton and R.C. Lewontin. 1978. Extent of genetic variation at a highly polymorphic esterase locus in *Drosophila pseudoobscura*. Proc. Natl. Acad. Sci. USA 75:5090–5093.

Coyne, J.A. and R. Lande. 1985. The genetic basis of species differences in plants. Am. Nat. 126:141–145.

Coyne, J.A. and M. Kreitman. 1986. Evolutionary genetics of two sibling species, *Drosophila simulans* and *D. sechellia*. Evolution 40:673–691.

Crick, F. 1979. Split genes and RNA splicing. Science 204:264–271.

Cross, S.R.H. and A.J. Birley. 1986. Restriction endonuclease map variation in the *Adh* region in populations of *Drosophila melanogaster*. Biochem. Genet. 24:415–433.

Crow, J.F. 1985. The neutrality–selection controversy in the history of evolution and population genetics. pp. 1–18. *In* T. Ohta and K. Aoki (eds.), *Population Genetics and Molecular Evolution*. Japanese Scientific Societies Press, Tokyo.

Crow, J.F. 1986. *Basic Concepts in Population, Quantitative, and Evolutionary Genetics*. W.H. Freeman, San Francisco.

Crow, J.F. and K. Aoki. 1982. Group selection for a polygenic behavioral trait: A differential proliferation model. Proc. Natl. Acad. Sci. USA 79:2628–2631.

Crow, J.F. and K. Aoki. 1984. Group selection for a polygenic behavioral trait: Estimating the degree of population subdivision. Proc. Natl. Acad. Sci. USA 81:6073–6077.

Crow, J.F. and C. Denniston (eds.). 1974. *Genetic Distance*. Plenum Press, New York.

Crow, J.F. and C. Denniston. 1988. Inbreeding and variance effective numbers. Evolution 42:482–495.

Crow, J.F. and M. Kimura. 1970. *An Introduction to Population Genetic Theory*. Harper & Row, New York.

Crow, J.F. and A.P. Mange. 1965. Measurement of inbreeding from the frequency of marriages between persons of the same surname. Eugen. Quart. 12:199–203.

Crow, J.F. and R.G. Temin. 1964. Evidence for the partial dominance of recessive lethal genes in natural populations of *Drosophila*. Am. Nat. 98:21–33.

Crozier, R.H. 1977. Evolutionary genetics of the Hymenoptera. Annu. Rev. Entomol. 22:263–288.

Crumpacker, D.W. 1967. Genetic loads in maize (*Zea mays*) and other cross-fertilized plants and animals. Evol. Biol. 1:306–424.

Curie-Cohen, M. 1982. Estimates of inbreeding in a natural population: A comparison of sampling properties. Genetics 100:339–358.

Curtis, S.E. and M.T. Clegg. 1984. Molecular evolution of chloroplast DNA sequences. Mol. Biol. Evol. 1:291–301.

Curtsinger, J.W. 1980. On the opportunity for polymorphism with sex linkage or haplodiploidy. Genetics 96:995–1006.

Curtsinger, J.W. 1984. Evolutionary landscapes for complex selection. Evolution 38:359–367.

Curtsinger, J.W. and M.W. Feldman. 1980. Experimental and theoretical analysis of the "sex ratio" polymorphism in *Drosophila pseudoobscura*. Genetics 94:445–466.

Daniels, G.R. and P.L. Deininger. 1985. Integration site preferences of the *Alu* family and similar repetitive DNA sequences. Nucleic Acids Res. 13:8939–8954.

Daniels, S.B., L.D. Strausbaugh, L. Ehrman and R. Armstrong. 1984. Sequences homologous to *P* elements occur in *Drosophila paulistorum*. Proc. Natl. Acad. Sci. USA 81:6794–6797.

Darnell, J., H. Lodish and D. Baltimore. 1986. *Molecular Cell Biology*. W.H. Freeman, New York.

Darwin, C. 1859. *On the Origin of Species by Means of Natural Selection*. Murray, London.

Darwin, C. 1871. *The Descent of Man and Selection in Relation to Sex*. D. Appleton and Co., New York.

Davies, J. and D.I. Smith. 1978. Plasmid-determined resistance to anti-microbial agents. Annu. Rev. Microbiol. 32:469–518.

Davis, P.S., M.W. Shen and B.H. Judd. 1987. Asymmetrical pairings of transposons in and proximal to the *white* locus of *Drosophila* account for four classes of regularly occurring exchange products. Proc. Natl. Acad. Sci. USA 84:174–178.

Dawood, M.M. and M.W. Strickberger. 1969. The effect of larval interaction on viability in *Drosophila melanogaster*. III. Effects of biotic residues. Genetics 63:213–220.

Dean, A.M., D.E. Dykhuizen and D.L. Hartl. 1986. Fitness as a function of β-galactosidase activity in *Escherichia coli*. Genet. Res. 48:1–8.

DeBenedictis, P.A. 1977. Studies in the dynamics of genetically variable populations. I. Frequency- and density-dependent selection in experimental populations of *Drosophila melanogaster*. Genetics 87:343–356.

De Jong, G. and W. Scharloo. 1976. Environmental determination of selective significance or neutrality of amylase variants in *Drosophila melanogaster*. Genetics 84:77–94.

Delarbre, C., Y. Kashi, P. Boursot, J.S. Beckmann, P. Kourilsky, F. Bonhomme and G. Gachelin. 1988. Phylogenetic distribution in the genus *Mus* of *t*-

complex specific DNA and protein markers: Inferences on the origin of *t*-haplotypes. Mol. Biol. Evol. 5:120–133.

Demers, G.W., K. Brech and R.C. Hardison. 1986. Long interspersed L1 repeats in rabbit DNA are homologous to L1 repeats of rodents and primates in an open reading frame region. Mol. Biol. Evol. 3:179–190.

Dempster, A.P., N.M. Laird and D.B. Rubin. 1977. Maximum likelihood estimation from incomplete data via the EM algorithm. J. Roy. Statist. Soc. B 39:1–38.

Dempster, E.R. 1955. Maintenance of genetic heterogeneity. Cold Spring Harbor Symp. Quant. Biol. 20:25–32.

Densmore, L.D., J.W. Wright and W.M. Brown. 1985. Length variation and heteroplasmy are frequent in mitochondrial DNA from parthenogenetic and bisexual lizards (genus *Cnemidophorus*). Genetics 110:689–707.

DeSalle, R., J. Slightom and E. Zimmer. 1986. The molecular through ecological genetics of abnormal abdomen. II. Ribosomal DNA polymorphism is associated with the abnormal abdomen syndrome in *Drosophila mercatorum*. Genetics 112:861–875.

DeSalle, R. and L.V. Giddings, 1986 Discordance of nuclear and mitochondrial DNA phylogenies in Hawaiian *Drosophila*. Proc. Natl. Acad. Sci. USA 83:6902–6906.

Dice, L.R. and W.B. Howard. 1951. Distance of dispersal by prairie deer mice from birthplaces to breeding sites. Contr. Lab. Vert. Biol. Univ. Mich. 50:1–15.

Dickerson, R.E. and I. Geis. 1983. *Hemoglobin: Structure, Function, Evolution and Pathology*. Benjamin/Cummings, Menlo Park, CA.

DiLella, A.G., J. Marvit, K. Brayton and S.L.C. Woo. 1987. An amino acid substitution involved in phenylketonuria is in linkage disequilibrium with DNA haplotype 2. Nature 327:333–336.

DiLella, A.G., J. Marvit, A.S. Lidsky, F. Gutler and S.L.C. Woo. 1986. Tight linkage between a splicing mutation and a specific DNA haplotype in phenylketonuria. Nature 322:799–803.

DiMichele, L. and D.A. Powers. 1982. Physiological basis for swimming endurance differences between LDH-B genotypes of *Fundulus heteroclitus*. Science 216:1014–1016.

Dobzhansky, Th. 1936. Studies in hybrid sterility. II. Localization of sterility factors in *Drosophila pseudoobscura* hybrids. Genetics 21:113–135.

Dobzhansky, Th. 1943. Genetics of natural populations. IX. Temporal changes in the composition of populations of *Drosophila pseudoobscura*. Genetics 28:162–186.

Dobzhansky, Th. 1970. *Genetics of the Evolutionary Process*. Columbia University Press, New York.

Dobzhansky, Th. and B. Spassky. 1944. Genetics of natural populations. XI. Manifestation of genetic variants of *Drosophila pseudoobscura* in different environments. Genetics 20:270–290.

Dobzhansky, Th., B. Spassky and T. Tidwell. 1963. Genetics of natural populations. XXXIV. Adaptive norm, genetic load, and genetic elite in *Drosophila pseudoobscura*. Genetics 48:1467–1485.

Donis-Keller, H. and 32 other authors. 1987. A genetic linkage map of the human genome. Cell 51:319–337.

Doolittle, F.W. and C. Sapienza. 1980. Selfish genes, the phenotype paradigm and genome evolution. Nature 284:601–603.

Doolittle, R.F. 1985. The genealogy of some recently evolved vertebrate proteins. Trends in Biochem. Sci. 10:233–237.

Dover, G. 1982. Molecular drive: A cohesive mode of species evolution. Nature 299:111–117.

Dudley, J.W. 1977. 76 generations of selection for oil and protein percentage in maize. pp. 459–473. *In* E. Pollak, O. Kempthorne, and T.B. Bailey Jr. (eds.), *International Conference on Quantitative Genetics*. Iowa State University Press, Ames, Iowa.

DuMouchel, W.H. and W.W. Anderson. 1968. The analysis of selection in experimental populations. Genetics 58:435–449.

Dutton, F.L. and H.M. Krider. 1985. Expression and amplification of the genes for ribosomal RNA in *bobbed* mutants of *Drosophila melanogaster*. Genet. Res. 45:155–165.

Dykhuizen, D.E., A.M. Dean and D.L. Hartl. 1987. Metabolic flux and fitness. Genetics 115:25–31.

Dykhuizen, D.E., J. de Framond and D.L. Hartl. 1984a. Selective neutrality of glucose-6-phosphate dehydrogenase allozymes in *Escherichia coli*. Mol. Biol. Evol. 1:162–170.

Dykhuizen, D.E., J. de Framond and D.L. Hartl. 1984b. Potential for hitchhiking in the *eda-edd-zwf* gene cluster in *Escherichia coli*. Genet. Res. 43:229–239.

Dykhuizen, D. and D.L. Hartl. 1980. Selective neutrality of 6PGD allozymes in *E. coli* and the effects of genetic background. Genetics 96:801–817.

Dykhuizen, D.E. and D.L. Hartl. 1983a. Selection in chemostats. Microb. Rev. 47:150–168.

Dykhuizen, D.E. and D.L. Hartl. 1983b. Functional effects of PGI allozymes in *Escherichia coli*. Genetics 105:1–18

Eanes, W.F., J. Hey and D. Houle. 1985a. Homozygous and hemizygous viability variation on the X

chromosome of *Drosophila melanogaster*. Genetics 111:831–844.

Eanes, W.F., B. Bingham, J. Hey and D. Houle. 1985b. Targeted selection experiments and enzyme polymorphism: Negative evidence for octanoate selection at the *G6PD* locus in *Drosophila melanogaster*. Genetics 109:379–391.

East, E.M. and A.J. Mangelsdorf. 1925. A new interpretation of the hereditary behavior of self-sterile plants. Proc. Natl. Acad. Sci. USA 11:166–171.

Easteal, S. 1985. Generation time and the rate of molecular evolution. Mol. Biol. Evol. 2:450–453.

Easterby, J.S. 1973. Coupled enzyme assays: A general expression for the transient. Biochim. Biophys. Acta 293:552–558.

Edwards, A.W.F. 1971. Distance between populations on the basis of gene frequencies. Biometrics 27:873–881.

Edwards, M.D., C.W. Stuber and J.F. Wendel. 1988. Molecular-marker facilitated investigations of quantitative trait loci in maize. I. Numbers, genomic distribution and types of gene action. Genetics 116:113–125.

Eggleston, W.B., D.M. Johnson-Schlitz and W.R. Engels. 1988. *P–M* dysgensis does not mobilize other transposable element families in *D. melanogaster*. Nature 331:368–370.

Ehrlich, P.R. and P.H. Raven. 1964. Butterflies and plants: A study in coevolution. Evolution 18:586–608.

Ehrman, L. 1970. The mating advantage of rare males in *Drosophila*. Proc. Natl. Acad. Sci. USA 65:345–348.

Ellstrand, N.C. 1984. Multiple paternity within the fruits of the wild radish *Raphanus sativus*. Am. Nat. 123:819–828.

Elton, C. and M. Nicholson. 1942. The ten-year cycle of the lynx in Canada. J. Animal Ecol. 11:215–244.

Enfield, F.D. 1980. Long term effects of selection: The limits to response. pp. 69–86. *In* A. Robertson (ed.), *Selection Experiments in Laboratory and Domestic Animals*. Commonwealth Agricultural Bureau, Slough, England.

Engelke, D.R., P.A. Hoener and F.S. Collins. 1988. Direct sequencing of enzymatically amplified human genomic DNA. Proc. Natl. Acad. Sci. USA 85:544–548.

Engels, W.R. 1981. Estimating genetic divergence and genetic variability with restriction endonucleases. Proc. Natl. Acad. Sci. USA 78:6329–6333.

Engels, W.R. 1983. The *P* family of transposable elements in *Drosophila*. Annu. Rev. Genet. 17:315–344.

Engels, W.R. 1988. *P* elements in *Drosophila*. *In* D. Berg and M. Hopew (eds.), *Mobile DNA*. ASM Publications.

Eshel, I. 1975. Selection on sex ratio and the evolution of sex determination. Heredity 34:351–361.

Eshel, I. 1977. On the founder effect and the evolution of altruistic traits: An ecological approach. Theor. Pop. Biol. 11:410–424.

Eshel, I. and M.W. Feldman. 1982. On evolutionary genetic stability of the sex ratio. Theor. Pop. Biol. 21:430–439.

Ewens, W.J. 1969. Mean fitness increases when fitnesses are additive. Nature 221:1076.

Ewens, W.J. 1972. The sampling theory of selectively neutral alleles. Theor. Pop. Biol. 3:87–112.

Ewens, W.J. 1979. *Mathematical Population Genetics*. Springer-Verlag, Berlin.

Ewens, W.J. 1982. On the concept of the effective population size. Theor. Pop. Biol. 21:373–378.

Ewens, W.J., R.S. Spielman and H. Harris. 1981. Estimation of genetic variation at the DNA level from restriction endonuclease data. Proc. Natl. Acad. Sci. USA 78:3748–3750.

Ewens, W.J. and G. Thomson. 1977. Properties of equilibria in multi-locus genetic systems. Genetics 87:807–819.

Ewer, J., M. Rosbach and J.C. Hall. 1988. An inducible promoter fused to the period gene in *Drosophila* conditionally rescues adult *per*-mutant arrhythmicity. Nature 333:82–84.

Falconer, D.S. 1955. Patterns of response in selection experiments with mice. Cold Spring Harbor Symp. Quant. Biol. 20:178–196.

Falconer, D.S. 197. Some results of the Edinburgh selection experiments with mice. pp. 101–115. *In* E. Pollak, O. Kempthorne and T.B. Bailey Jr. (eds.), *International Conference on Quantitative Genetics*, Iowa State University Press, Ames.

Falconer, D.S. 1981. *Introduction to Quantitative Genetics*, 2nd Ed. Longman, London.

Falconer, D.S. 1985. A note on Fisher's "average effect" and "average excess." Genet. Res. 46:337–347.

Falk, C.T. and C.C. Li. 1969. Negative assortative mating: Exact solution to a simple model. Genetics 62:215–223.

Fedoroff, N.V. 1984. Transposable genetic elements in maize. Sci. Amer. (June):85–98.

Feldman, M.W. 1972. Selection for linkage modification I. Random mating populations. Theor. Pop. Biol. 3:324–346.

Feldman, M.W. and F.B. Christiansen. 1975. The effect of population subdivision on two loci without selection. Genet. Res. 24:151–162.

Feldman, M.W., F.B. Christiansen and L.D. Brooks. 1980. Evolution of recombination in a constant environment. Proc. Natl. Acad. Sci. USA 77:4838–4841.

Feldman, M.W., F.B. Christiansen and U. Liberman. 1983. On some models of fertility selection. Genetics 105:1003–1010.

Feldman, M.W., I. Franklin and G. Thomson. 1974. Selection in complex genetic systems. I. The symmetric equilibria of the three-locus symmetric viability model. Genetics 76:135–162.

Feldman, M.W. and R.C. Lewontin. 1975. The heritability hang-up. Science 190:1163–1168.

Feldman, M.W., R.C. Lewontin, I.R. Franklin and F.B. Christiansen. 1975. Selection in complex genetic systems. III. An effect of allele multiplicity with two loci. Genetics 79:333–347.

Feldman, M.W. and U. Liberman. 1985. A symmetric two-locus fertility model. Genetics 109:229–253.

Feldman, M.W. and U. Liberman. 1986. An evolutionary reduction principle for genetic modifiers. Proc. Natl. Acad. Sci. USA 83:4824–4827.

Feldman, M.W., M. Nabholz and W.F. Bodmer. 1969. Evolution of the Rh polymorphism: A model for the interaction of incompatibility, reproductive compensation and heterozygote advantage. Am. J. Hum. Genet. 21:171–193.

Feldman, M.W. and E.A.C. Thomas. 1987. Behavior-dependent contexts for repeated plays of the prisoner's dilemma II. Dynamical aspects of the evolution of cooperation. J. Theor. Biol. 128:297–315.

Felsenstein, J. 1971. Inbreeding and variance effective numbers in populations with overlapping generations. Genetics 68:581–597.

Felsenstein, J. 1974. The evolutionary advantage of recombination. Genetics 78:737–756.

Felsenstein, J. 1978. Cases in which parsimony or compatibility methods will be positively misleading. Syst. Zool. 27:401–410.

Felsenstein, J. 1981. Evolutionary trees from DNA sequences: A maximum likelihood approach. J. Mol. Evol. 17:368–376.

Felsenstein, J. 1982. Numerical methods for inferring evolutionary trees. Quart. Rev. Biol. 57:379–404.

Ferris, S.D., R.D. Sage, C.M. Huang, J.T. Nielsen, U. Ritte and A.C. Wilson. 1983a. Flow of mitochondrial DNA across a species boundary. Proc. Natl. Acad. Sci. USA 80:2290–2294.

Ferris, S.D., R.D. Sage, E.M. Prager, U. Ritte and A.C. Wilson. 1983b. Mitochondrial DNA evolution in mice. Genetics 105:681–721.

Finney, D. J. 1952. The equilibrium of a self-incompatible polymorphic species. Genetica 26:33.

Fisher, R.A. 1918. The correlations between relatives on the supposition of Mendelian inheritance. Trans. Roy. Soc. Edinburgh 52:399–433.

Fisher, R.A. 1922. On the dominance ratio. Proc. Roy. Soc. Edin. 42:321–431.

Fisher, R.A. 1930. *The Genetical Theory of Natural Selection*, 1st Ed. Clarendon, Oxford.

Fisher, R.A. 1958. *The Genetical Theory of Natural Selection*, 2nd Ed. Dover Press, New York.

Fitch, W.M. 1977. On the problem of discovering the most parsimonious tree. Am. Nat. 3:223–257.

Fitch, W.M. and E. Margoliash. 1967. Construction of phylogenetic trees. Science 155:279–284.

Flint, J., A.V.S. Hill, D.K. Bowden, S.J. Oppenheimer, P.R. Sill, S.W. Sereantson, J. Bana-Koiri, K. Bhatia, M.P. Alpers, A.J. Boyce, D.J. Weatherall and J.B. Clegg. 1986. High frequencies of α-thalassemia are the result of natural selection by malaria. Nature 321:744–750.

Foley, P. 1987. Molecular clock at loci under stabilizing selection. Proc. Natl. Acad. Sci. USA 84:7996–8000.

Ford, H.A. 1971. The degree of mimetic protection gained by new partial mimics. Heredity 27:227–236.

Foster, G.G., M.J. Whitten, T. Prout and R. Gill. 1972. Chromosome rearrangements for the control of insect pests. Science 176:875–880.

Frankham, R., D.A. Briscoe and R.K. Nurthen. 1978. Unequal crossing over at the rRNA locus as a source of quantitative genetic variation. Nature 272:80–81.

Frankham, R., D.A. Briscoe, R.K. Nurhen. 1980. Unequal crossing over at the rRNA tandem as a source of quantitative genetic variation in *Drosophila*. Genetics 95:727–742.

Franklin, I.R. and M.W. Feldman. 1977. Two loci with two alleles: Linkage equilibrium and linkage disequilibrium can be simultaneously stable. Theor. Pop. Biol. 12:95–113.

Franklin, I. and R.C. Lewontin. 1970. Is the gene the unit of selection? Genetics 65:707–734.

Frischauf, A.-M. 1985. The *T/t* complex of the mouse. Trends in Genet. 1:100–103.

Frydenberg, O. 1963. Population studies of a lethal mutant in *Drosophila melanogaster*. I. Behavior in populations with discrete generations. Hereditas 50:89–116.

Fryer, G. and T.D. Iles. 1972. *The Cichlid Fishes of the Great Lakes of Africa: Their Biology and Evolution*. T.F.H. Publications, Hong Kong.

Fuerst, P.A., R. Chakraborty and M. Nei. 1977. Statistical studies on protein polymorphism in natural populations. I. Distribution of single-locus heterozygosity. Genetics 86:455–483.

Fuerst, P.A. and R.E. Ferrell. 1980. The stepwise mutation model: An experimental evaluation utilizing hemoglobin variants. Genetics 94:185–201.

Fuerst, P.A. and T. Maruyama. 1986. Patterns of linkage disequilibrium generated by rapid changes in population size. Genetics 113:S44.

Futuyma, D.J. and M. Slatkin (eds.). 1983. *Coevolution*. Sinauer Associates, Sunderland, MA.

Gaffney, B. and E.P. Cunningham. 1988. Estimation of genetic trend in racing performance of thoroughbred horses. Nature 332:722–724.

Gardner, P., D. H. Smith, H. Beer and R.C. Moellering, Jr. 1969. Recovery of resistance (R) factors from a drug-free community. Lancet (Oct. 11, 1969):774–776.

Gause, G.F. 1934. *The Struggle for Existence*. Williams and Wilkins, Baltimore.

Geer, B.W., S.W. McKechnie and M.L. Langevin. 1986. The effect of dietary ethanol on the composition of lipids of *Drosophila melanogaster* larvae. Biochem. Genet. 24:51–69.

George, J.T. and R.C. Elston. 1987. Testing the association between polymorphic markers and quantitative traits in pedigrees. Genet. Epidem. 4:193–200.

Gibbs, K.L. and P.R. Grant. 1987. Oscillating selection on Darwin's finches. Nature 327:511–513.

Giesel, J.T. 1979. Genetic co-variation of survivorship and other fitness indices in *Drosophila melanogaster*. Exp. Geront. 14:323–328.

Giesel, J.T. 1986. Genetic correlation structure of life history variables in outbred, wild *Drosophila melanogaster*: Effects of photoperiod regimen. Am. Nat. 128:593–603.

Giesel, J.T., P.A. Murphy and M.N. Manlove. 1982. The influence of temperature on genetic interrelationships of life history traits in a population of *Drosophila melanogaster*: What tangled data sets we weave. Am. Nat. 119:464–479.

Gilbert, W. 1986. The road not taken. Nature 320:485–486.

Gillespie, J.H. 1973. Natural selection with varying selection coefficients — a haploid model. Genet. Res. 21:115–120.

Gillespie, J.H. 1974. Polymorphism in patchy environments. Am. Nat. 108:145–151.

Gillespie, J.H. 1975. The role of migration in the genetic structure of populations in temporally and spatially varying environments. I. Conditions for polymorphism. Am. Nat. 109:127–135.

Gillespie, J.H. 1978. A general model to account for enzyme variation in natural populations V. The SAS-CFF model. Theor. Pop. Biol. 14:1–45.

Gillespie, J.H. 1981a. Mutation modification in a random environment. Evolution 35:468–476.

Gillespie, J.H. 1981b. Evolution of the mutation rate at a heterotic locus. Proc. Natl. Acad. Sci. USA 78:2452–2454.

Gillespie, J.H. 1984. The molecular clock may be an episodic clock. Proc. Natl. Acad. Sci. USA 81:8009–8013.

Gillespie, J.H. 1986a. Variability of evolutionary rates of DNA. Genetics 113:1077–1091.

Gillespie, J.H. 1986b. Natural selection and the molecular clock. Mol. Biol. Evol. 3:138–155.

Gillespie, J.H. 1988. More on the overdispersed molecular clock. Genetics 118:385–386.

Gillespie, J.H. and K. Kojima. 1968. The degree of polymorphism in enzymes involved in energy production compared to that in nonspecific enzymes in two *D. ananassae* populations. Proc. Natl. Acad. Sci. USA 61:582–585.

Gillespie, J.H. and C. H. Langley. 1974. A general model to account for enzyme variation in natural populations. Genetics 76:837–848.

Gillings, M.R., R. Frankham, J. Speirs and M. Whaley. 1987. X-Y exchange and coevolution of the X and Y rDNA arrays in *Drosophila melanogaster*. Genetics 116:241–251.

Gingerich, P.D. 1986. Temporal scaling of molecular evolution in primates and other mammals. Mol. Biol. Evol. 3:205–221.

Ginzburg, L.R. and E.M. Golenberg. 1985. *Lectures in Theoretical Population Biology*. Prentice-Hall, Englewood Cliffs, NJ.

Gliddon, C. and C. Strobeck. 1975. Necessary and sufficient conditions for multiple-niche polymorphisms in haploids. Am. Nat. 109:233–235.

Glover, D.M. and D.S. Hogness. 1977. A novel arrangement of the 18S and 28S sequences in a repeating unit of *Drosophila melanogaster* rDNA. Cell 10:167–176.

Golding, G.B. 1984. The sampling distribution of linkage disequilibrium. Genetics 108:257–274.

Golding, G.B. 1987. Nonrandom patterns of mutation are reflected in evolutionary divergence and may cause some of the unusual patterns observed in sequences. pp. 151–172. *In* V. Loeschcke (ed.), *Genetic Constraints on Adaptive Evolution*. Springer-Verlag, Berlin.

Golding, G.B., C.F. Aquadro and C.H. Langley. 1986. Sequence evolution within populations under multiple types of mutation. Proc. Natl. Acad. Sci. USA 83:427–431.

Goodman, M., B.F. Koop, J. Czelusniak and M.L. Weiss. 1984. The η-globin gene: Its long evolutionary history in the β-globin gene family of mammals. J. Mol. Biol. 180:803–823.

Gotoh, O., J.-I. Hayashi, H. Yonekawa and Y. Tagashira. 1979. An improved method for estimating sequence divergence between related DNAs from changes in restriction endonuclease cleavage sites. J. Mol. Evol. 14:301–310.

Gould, S.J. and R.C. Lewontin. 1979. The spandrels of San Marco and the Panglossian paradigm. A critique of the adaptationist programme. Proc. Roy. Soc. B. 205:581–598.

Gould, S.J., N.L. Gilinsky and R.Z. German. 1987. Asymmetry of lineages and the direction of evolutionary time. Science 236:1437–1441.

Gray, P.W. and D.V. Goeddel. 1983. Cloning and expression of murine immune interferon cDNA. Proc. Natl. Acad. Sci. USA 80:5842–5846.

Greaves, J.H., R. Redfern, P.B. Ayres and J.E. Gill. 1977. Warfarin resistance: A balanced polymorphism in the Norway rat. Genet. Res. 30:257–263.

Griffiths, D.J. 1950. The liability of seed crops of perennial rye-grass (*Lolium perenne*) to contamination by wind-borne pollen. J. Agric. Res. 40:19–38.

Grun, P. 1976. *Cytoplasmic Genetics and Evolution*. Columbia University Press, New York.

Gupta, A.P. and R.C. Lewontin. 1982. A study of reaction norms in natural populations of *Drosophila pseudoobscura*. Evolution 36:934–948.

Gusella, J.F., N.S. Wexler, P.M. Conneally, S.L. Naylor, M.A. Anderson, M.A. Tanzi, P.C. Watkins, K. Ottina, M.R. Wallace, A.Y. Sakaguchi, A.B. Young, I. Shouson, E. Bonilla and J.B. Martin. 1983. A polymorphic DNA marker genetically linked to Huntington's disease. Nature 306:234–238.

Gwynne, D.T. 1984. Sexual selection and sexual differences in Mormon crickets (Orthoptera: Tettigoniidae, *Anabrus simplex*). Evolution 38:1011–1022.

Gwynne, D.T. 1988. Courtship feeding and the fitness of female katydids (Orthoptera: Tettigoniidae). Evolution 42:545–555.

Hadeler, K.P and U. Liberman. 1975. Selection models with fertility differences. J. Math. Biol. 2:19–32.

Hairston, N.G. and W.E. Walton. 1986. Rapid evolution of a life history trait. Proc. Natl. Acad. Sci. USA 83:4831–4833.

Haldane, J.B.S. 1927. A mathematical theory of natural and artificial selection. Part IV. Proc. Camb. Phil. Soc. 23:607–615.

Haldane, J.B.S. 1932. *The Causes of Evolution*. Longman, New York.

Haldane, J.B.S. 1949. The association of characters as a result of inbreeding and linkage. Ann. Eugen. 14:288–292.

Haldane, J.B.S. 1956. The estimation of viabilities. J. Genet. 54:294–296.

Haldane, J.B.S. and S.D. Jayakar. 1963. Polymorphism due to selection of varying direction. J. Genet. 58:237–242.

Haldane, J.B.S. and S.D. Jayakar. 1964. Equilibria under natural selection at a sex-linked locus. J. Genet. 59:29–36.

Hale, L.R. and R.S. Singh. 1986. Extensive variation and heteroplasmy in size of mitochondrial DNA among geographic populations of *Drosophila melanogaster*. Proc. Natl. Acad. Sci. USA 83:8813–8817.

Hall, B.G. 1983. Evolution of new metabolic functions in laboratory organisms. pp. 234–257. *In* M. Nei and R. Koehn (eds.), *Evolution of Genes and Proteins*. Sinauer Associates, Sunderland, MA.

Hall, B.G., S. Yokoyama and D.H. Calhoun. 1983. Role of cryptic genes in microbial evolution. Mol. Biol. Evol. 1:109–124.

Hall, B.G. and P.W. Betts. 1987. Cryptic genes for cellobiose utilization in natural isolates of *Escherichia coli*. Genetics 115:431–439.

Hall, B.G., S. Yokoyama and D.H. Calhoun. 1983. Role of cryptic genes in microbial evolution., Mol. Biol. Evol. 1:109–124.

Hamilton, W.D. 1964. The genetical evolution of social behavior. J. Theor. Biol. 7:1–16, and 17–52.

Hamilton, W.D. 1966. The moulding of senescence by natural selection. J. Theor. Biol. 12:12–45.

Hamilton, W.D. 1967. Extraordinary sex ratios. Science 156:477–488.

Hamilton, W.D. 1972. Altruism and related phenomena, mainly in social insects. Annu. Rev. Ecol. Syst. 3:193–232.

Hamrick, J.L. and R.W. Allard. 1972. Microgeographical variation in allozyme frequencies in *Avena barbata*. Proc. Natl. Acad. Sci. USA 69:2100–2104.

Hardies, S.C., M.H. Edgell and C.A. Hutchinson III. 1984. Evolution of the mammalian β-globin gene cluster. J. Biol. Chem. 259:3748–3756.

Hardies, S.C., S.L. Martin, C.F. Voliva, C.A. Hutchison III and M.H. Edgell. 1986. An analysis of replacement and synonymous changes in the rodent L1 repeat family. Mol. Biol. Evol. 3:109–125.

Hardin, G. 1960. The competitive exclusion principle. Science 131:1292–1298.

Hardison, R.C. 1984. Comparison of the β-like globin gene families of rabbits and humans indicates that the gene cluster 5′-η-γ-δ-β-3′ predates the mammalian radiation. Mol. Biol. Evol. 1:390–410.

Hardison, R.C. and J.B. Margot. 1984. Rabbit globin

pseudogene ψβ2 is a hybrid of δ- and β-globin gene sequences. Mol. Biol. Evol. 1:302–316.
Hardison, R.C. and R. Printz. 1985. Variability within the rabbit C repeat and sequences shared with other SINES. Nuc. Acids Res. 13:1073–1088.
Hardy, G.H. 1908. Mendelian proportions in a mixed population. Science 28:41–50. [Reprinted in Peters, J.H. (ed.). 1959. *Classic Papers in Genetics*. Prentice-Hall, Englewood Cliffs, NJ.]
Harris, H. 1966. Enzyme polymorphisms in man. Proc. Roy. Soc. B. 164:298–310.
Harris, H. 1969. Enzyme and protein polymorphism in human populations. Brit. Med. Bull. 25:5–13.
Harris, H. and D.A. Hopkinson. 1972. Average heterozygosity in man. Ann. Hum. Genet. 36:9–20.
Harris, H. and D.A. Hopkinson. 1976. *Handbook of Enzyme Electrophoresis in Human Genetics*. American Elsevier, New York.
Harris, H., D.A. Hopkinson and Y.H. Edwards. 1977. Polymorphism and the subunit structure of enzymes. A contribution to the neutralist-selectionist controversy. Proc. Natl. Acad. Sci. USA 74:698–701.
Harris, S., J.R. Thackeray, A.J. Jeffreys and M.L. Weiss. 1986. Nucleotide sequence analysis of the lemur β-globin gene family: Evidence for major rate fluctuations in globin polypeptide evolution. Mol. Biol. Evol. 3:465–484.
Harris, T., E.F. Cook, R. Garrison, et al. 1988. Body mass index and mortality among nonsmoking older persons: The Framingham Heart Study. J. Am. Med. Assoc. 259:1520–1524.
Harrison, R.G. 1983. Barriers to gene exchange between closely related cricket species. I. Laboratory hybridization studies. Evolution 37:245–251.
Harrison, R.G. 1985. Barriers to gene exchange between closely related cricket species. II. Life cycle variation and temporal isolation. Evolution 39:244–259.
Harrison, R.G., D.M. Rand and W.C. Wheeler. 1985. Mitochondrial DNA size variation within individual crickets. Science 228:1446–1447.
Harrison, R.G., D.M. Rand and W.C. Wheeler. 1987. Mitochondrial DNA variation in field crickets across a narrow hybrid zone. Mol. Biol. Evol. 4:144–158.
Hartl, D.L. 1972. A fundamental theorem of natural selection for sex linkage or arrhenotoky. Am. Nat. 106:516–524.
Hartl, D.L. 1975. Modifier theory and meiotic drive. Theor. Pop. Biol. 7:168–174.
Hartl, D.L. 1980. Genetic dissection of segregation distortion. III. Unequal recovery of reciprocal recombinants. Genetics 96:685–696.
Hartl, D.L. 1988. *A Primer of Population Genetics*, 2nd Ed. Sinauer Associates, Inc., Sunderland, MA.
Hartl, D.L. and S.W. Brown. 1970. The origin of male haploid genetic systems and their expected sex ratio. Theor. Pop. Biol. 1:165–190.
Hartl, D.L. and R.D. Cook. 1973. Balanced polymorphisms of quasi-neutral alleles. Theor. Pop. Biol. 4:163–172.
Hartl, D.L., A.M. Dean and D.E. Dykhuizen. 1986. The molecular biology of natural selection: Reply to Burton and Place. Genetics 114:1037–1039.
Hartl, D.L. and D.E. Dykhuizen. 1981. Potential for selection among nearly neutral allozymes of 6-phosphogluconate dehydrogenase in *Escherichia coli*. Proc. Natl. Acad. Sci. USA 78:6344–6348.
Hartl, D.L. and D.E. Dykhuizen. 1984. The population genetics of *Escherichia coli*. Annu. Rev. Genet. 18:31–68.
Hartl, D.L., D.E. Dykhuizen and A.M. Dean. 1985. Limits of adaptation: The evolution of selective neutrality. Genetics 111:655–674.
Hartl, D.L., D.E. Dykhuizen, R.D. Miller, L. Green and J. de Framond. 1983. Transposable element IS50 improves growth rate of *E. coli* cells without transposition. Cell 35:503–510.
Hartl, D.L and Y. Hiraizumi. 1976. Segregation distortion. pp. 615–666. *In* M. Ashburner and E. Novitski (eds.), *Genetics and Biology of Drosophila*, Vol. 1b. Academic Press, New York.
Hartl, D.L. and S.A. Sawyer. 1988. Why do unrelated insertion sequences occur together in the genome of *Escherichia coli*? Genetics 118:537–541.
Hastings, A. 1982. Unexpected behavior in two-locus genetic systems: An analysis of marginal underdominance at a stable equilibrium. Genetics 102:129–138.
Hastings, A. 1984. Linkage disequilibrium, selection and recombination at three loci. Genetics 106:153–164.
Hastings, A. 1985a. Four simultaneously stable polymorphic equilibria in two-locus two-allele models. Genetics 109:255–261.
Hastings, A. 1985b. Multilocus population genetics with weak epistasis. I. Equilibrium properties of two-locus two-allele models. Genetics 109:799–812.
Hastings, A. 1986a. Multilocus population genetics with weak epistasis. II. Equilibrium properties of multilocus models: What is the unit of selection? Genetics 112:157–171.
Hastings, A. 1986b. Limits to the relationship among recombination, disequilibrium and epistasis in two-locus models. Genetics 113: 177–185.
Hastings, A. 1987. Substitution rates under stabilizing selection. Genetics 116:479–486.

Hauswirth, W.W. and P.J. Laipis. 1982. Mitochondrial DNA polymorphism in a maternal lineage of Holstein cows. Proc. Natl. Acad. Sci. USA 79:4686–4690.

Hayasaka, K., S. Horai, T. Gojobori, T. Shotake, K. Nozawa and E. Matsunaga. 1988. Phylogenetic relationships among Japanese, rhesus, Formosan, and crab-eating monkeys, inferred from restriction-enzyme analysis of mitochondrial DNAs. Mol. Biol. Evol. 5:270–281.

Hayashida, H., H. Toh, R. Kikuno and T. Miyata. 1985. Evolution of influenza virus genes. Mol. Biol. Evol. 2:289–303.

Haymer, D.S. and D.L Hartl. 1981. Using frequency distributions to detect selection: Inversion polymorphisms in *Drosophila pseudoobscura*. Evolution 35:1243–1246.

Haymer, D.S. and D.L. Hartl. 1983a. The experimental assessment of fitness in *Drosophila melanogaster*. I. Comparative measures of competitive reproductive success. Genetics 102:455–466.

Haymer, D.S. and D.L Hartl. 1983b. The experimental assessment of fitness in *Drosophila*. II. A comparison of competitive and noncompetitive measures. Genetics 104:343–352.

Head, G., R.M. May and L. Pendleton. 1987. Environmental determination of sex in reptiles. Nature 329:198–199.

Hedrick, P.W. 1972. Maintenance of genetic variation with a frequency-dependent selection model as compared to the overdominant model. Genetics 72:771–775.

Hedrick, P.W. 1975. Genetic similarity and distance: Comments and comparisons. Evolution 29:362–366.

Hedrick, P.W. 1976. Simulation of X-linked selection in *Drosophila*. Genetics 83:551–571.

Hedrick, P.W. 1987. Gametic disequilibrium measures: Proceed with caution. Genetics 117:331–341.

Hedrick, P.W., S. Jain and L. Holden. 1978. Multilocus systems in evolution. Evol. Biol. 11:101–184.

Hedrick, P.W. and E. Murray. 1983. Selection and measurement of fitness. pp. 61–104. *In* M. Ashburner, H. Carson and J.N. Thompson (eds.), *The Genetics and Biology of Drosophila*, Vol. 3b. Academic Press, London.

Hedrick, P.W. and G. Thomson. 1983. Evidence for balancing selection at HLA. Genetics 104: 449–456.

Hedrick, P.W. and G. Thomson. 1988. Maternal–fetal interactions and the maintenance of HLA polymorphism. Genetics 119:205–212.

Hedrick, P.W., G. Thomson and W. Klitz. 1986. Evolutionary genetics: HLA as an exemplary system. pp. 583–606. *In* S. Karlin and E. Nevo (eds.), *Evolutionary Processes and Theory*. Academic Press, New York.

Heinrich, R. and S.M. Rapoport. 1983. The utility of mathematical models for the understanding of metabolic systems. Bioch. Soc. Trans. 11:31–35.

Henderson, N.R. and D.M. Lambert. 1982. No significant deviation from random mating of worldwide populations of *Drosophila melanogaster*. Nature 300:437–440.

Herre, E.A. 1987. Optimality, plasticity and selective regime in fig wasp sex ratios. Nature 329:627–629.

Hickey, D.A. 1979. Selection on amylase allozymes in *Drosophila melanogaster*: Selection experiments using several independently derived pairs of chromosomes. Evolution 33:1128–1137.

Hickey, D.A. and B.F. Benkel. 1987. Regulation of amylase gene expression in *Drosophila melanogaster*: Amylase as a model experimental system. CRC Biot. 5:229–241.

Hilbish, T.J. and R.K. Koehn. 1985. The physiological basis of natural selection at the *LAP* locus. Evolution 39:1302–1317.

Hill, W.G. 1972. Effective size of populations with overlapping generations. Theor. Pop. Biol. 3:278–289.

Hill, W.G. 1974. Estimation of linkage disequilibrium in randomly mating populations. Heredity 33:229–239.

Hill, W.G. 1975. Tests for association of gene frequencies at several loci in random mating diploid populations. Biometrics 31:881–888.

Hill, W.G. and A. Robertson. 1968. Linkage disequilibrium in finite populations Theor. Appl Genet. 33:54–78.

Hinds, K.R. and G.W. Litman. 1986. Major reorganization of immunoglobulin V_H segments of elements during vertebrate evolution. Nature 320:546–549.

Hiraizumi, Y. 1961. Negative correlation between rate of development and female fertility in *Drosophila melanogaster*. Genetics 46:615–624.

Hiraizumi, Y. 1964. Prezygotic selection as a factor in the maintenance of variability. Cold Spring Harbor Symp. Quant. Biol. 29:51–60.

Hiraizumi, Y. 1984. Interaction in transmission frequency between the second and the third chromosomes in *Drosophila melanogaster*. Genetics 106:669–677.

Hiraizumi, Y. 1985. Genetics of factors affecting the life history of *Drosophila melanogaster*. I. Female productivity. Genetics 110:453–464.

Hiraizumi, Y. and J.F. Crow. 1960. Heterozygous ef-

fects on viability, fertility, rate of development and longevity of *Drosophila* chromosomes that are lethal when homozygous. Genetics 45:1071–1083.

Hiraizumi, Y. and A.M. Thomas. 1984. Suppressor systems of *Segregation Distorter* (*SD*) chromosomes in natural populations of *Drosophila melanogaster*. Genetics 106:279–292.

Hoekstra, R.F., R. Bijlsma and A.J. Dolman. 1985. Polymorphism from environmental heterogeneity: Models are only robust if the heterozygote is close in fitness to the favored homozygote in each environment. Genet. Res. 45:299–314.

Hoffman, A.A. and F.M. Cohan. 1987. Genetic divergence under uniform selection. III. Selection for knockdown resistance to ethanol in *Drosophila melanogaster* populations and their replicate lines. Heredity 58:425–433.

Hoffman, A.A. and M. Turelli. 1988. Unidirectional incompatibility in *Drosophila simulans*: Inheritance, geographic variation, and fitness effects. Genetics 119:435–444.

Hoffman, A.A., M. Turelli and G.M. Simmons. 1986. Unidirectional incompatibility between populations of *Drosophila simulans*. Evolution 40:692–701.

Hogg, J.T. 1984. Mating in Bighorn sheep: Multiple creative male strategies. Science 225:526–529.

Holgate, P. 1966. A mathematical study of the founder principle of evolutionary genetics. J. Appl. Prob. 3:115–128.

Holland, J., K. Spindler, F. Horodyski, E. Grabau, S. Nichol and S. VandePol. 1982. Rapid evolution of RNA genomes. Science 215:1577–1585.

Hollingsworth, T.H. 1969. *Historical Demography*. Cornell University Press, Ithaca, New York.

Holmquist, R., M.M. Miyamoto and M. Goodman. 1988. Analysis of higher-primate phylogeny from transversion differences in nuclear and mitochondrial DNA by Lake's methods of evolutionary parsimony and operator metrics. Mol. Biol. Evol. 5:217–236.

Holsinger, K.E. and M.W. Feldman. 1983a. Linkage modification with mixed random mating and selfing: A numerical study. Genetics 103:323–333.

Holsinger, K.E. and M.W. Feldman. 1983b. Modifiers of mutation rate: Evolutionary optimum with complete selfing. Proc. Natl. Acad. Sci. USA 80:6732–6734.

Holsinger, K.E., M.W. Feldman and L. Altenberg. 1986. Selection for increased mutation rates with fertility differences between matings. Genetics 112:909–922.

Holsinger, K.E., M.W. Feldman and F.B. Christiansen. 1984. The evolution of self-fertilization in plants: A population genetic model. Am. Nat. 124:446–453.

Hood, L., M. Kronenberg and T. Hunkapiller. 1985. T cell antigen receptors and the immunoglobulin supergene family. Cell 40:225–229.

Hubby, J.L. and R.C. Lewontin. 1966. A molecular approach to the study of genic heterozygosity in natural populations. I. The number of alleles at different loci in *Drosophila pseudoobscura*. Genetics 54:577–594.

Hudson, R.R. 1985. The sampling distribution of linkage disequilibrium under an infinite alleles model without selection. Genetics 109:611–631.

Hudson, R.R. 1987. Estimating the recombination parameter of a finite population without selection. Genet. Res. 50:245–250.

Hudson, R.R. and N.L. Kaplan. 1985. Statistical properties of the number of recombination events in the history of a sample of DNA sequences. Genetics 111:147–164.

Hudson, R.R., M. Kreitman and M. Aguade. 1987. A test of neutral molecular evolution based on nucleotide data. Genetics 116:153–159.

Huettel, M.D. and G.L. Bush. 1972. The genetics of host selection and its bearing on sympatric speciation in *Procecidochares* (Diptera: Tephritidae). Entomol. Exp. Appl. 15:465–480.

Hughes, D.M. and A.G. Clark. 1988. Analysis of the genetic structure of life history of *Drosophila melanogaster* using recombinant extracted lines. Evolution 42:1309–1320.

Hughes, A.L. and M. Nei. 1988. Pattern of nucleotide substitution at major histocompatibility complex class I loci reveals overdominant selection. Nature 355:167–170.

Hunkapiller, T. and L. Hood. 1986. The growing immunoglobulin supergene family. Nature 323:15–16.

Hwu, H.R., J.W. Roberts, E.H. Davidson and R.J. Britten. 1986. Insertion and/or deletion of many repeated DNA sequences in human and higher ape evolution. Proc. Natl. Acad. Sci. USA 83:3875–3879

Ikemura, T. 1985a. Codon usage and tRNA content in unicellular and multicellular organisms. Mol. Biol. Evol. 2:13–34.

Ikemura, T. 1985b. Codon usage, tRNA content, and rate of synonymous substitution. pp. 385–406. *In* T. Ohta and K. Aoki (eds.), *Population Genetics and Molecular Evolution*, Japanese Scientific Societies Press, Tokyo.

Ives, P.T. 1945. The genetic structure of American populations of *Drosophila melanogaster*. Genetics 30:167–196.

Jacob, F. 1977. Evolution and tinkering. Science 196:1161–1166.

Jacobson, J.W., M.M. Medhora and D.L. Hartl. 1988. Molecular structure of a somatically unstable transposable element in *Drosophila*. Proc. Natl. Acad. Sci. USA 83:8684–8688.

Jacquard, A. 1983. Heritability: One word, three concepts. Biometrics 39:465–477.

Jaenike, J. 1985a. Parasite pressure and the evolution of amanitin tolerance in *Drosophila*. Evolution 39:1295–1301.

Jaenike, J. 1985b. Genetic and environmental determinants of food preference in *Drosophila tripunctata*. Evolution 39:362–369.

Jaenike, J. 1988. Parasitism and male mating success in *Drosophila testacea*. Am. Nat. 131:774–780.

Jaenike, J., D.A. Grimaldi, A.E. Sluder and A.L. Greenleaf. 1983. α-amanitin tolerance in mycophagous *Drosophila*. Science 221:165–167.

Jain, S.K. and A.D. Bradshaw. 1966. Evolutionary divergence among adjacent plant populations. I. The evidence and its theoretical analysis. Heredity 21:407–441.

Jayakar, S.D. 1970. On the detection and estimation of linkage between a locus influencing a quantitative character and a marker locus. Biometrics 26:451–464.

Johannsen, W. 1903. Über Erblichkeit in Populationen und in reinen Linien. Gustav Fischer, Jena. [Translated in part by Peters, J.A. (ed.). 1959. *Classic Papers in Genetics*, pp. 21–26. Prentice-Hall, Englewood Cliffs, NJ.]

Johannsen, W. 1909. Elemente der exackten Erblichkeitslehre. Gustav Fischer, Jena.

Johnson, K.R., J.E. Wright and B. May. 1987. Linkage relationships reflecting ancestral tetraploidy in salmonid fish. Genetics 116:579–591.

Jones, J.S. 1982. Genetic differences in individual behavior associated with shell polymorphism in the snail *Cepaea nemoralis*. Nature 298:749–750.

Jones, J.S. 1987. An asymmetrical view of fitness. Nature 325:298–299.

Jukes, T.H. and C.R. Cantor. 1969. Evolution of protein molecules. pp. 21–132. *In* H.N. Munro (ed.), *Mammalian Protein Metabolism III*. Academic Press, New York.

Jungen, H. and D.L. Hartl. 1979. Average fitness of populations of *Drosophila melanogaster* as estimated using compound-autosome strains. Evolution 33:359–370.

Kacser, H. and J.A. Burns. 1973. The control of flux. Symp. Soc. Exp. Biol. 32:65–104.

Kacser, H. and J.A. Burns. 1981. The molecular basis of dominance. Genetics 97:639–666.

Kacser, H. and J.W. Porteous. 1987. Control of metabolism: What do we have to measure? Trends Biochem. Sci. 12:5–14.

Kallman, K.D. 1984. A new look at sex determination in poecilid fishes. pp. 95–171. *In* B.J. Turner (ed.), *Evolutionary Biology of Fishes*. Plenum, New York.

Kan, Y.W. and A.M. Dozy. 1978. Polymorphism of DNA sequence adjacent to the human β-globin structural gene: Relationship of sickle cell anemia. Proc. Natl. Acad. Sci. USA 75:5631–5635.

Kaplan, N. 1983. Statistical analysis of restriction enzyme map data and nucleotide sequence data. pp. 75–106. *In* B.S. Weir (ed.), *Statistical Analysis of DNA Sequence Data*. Marcel Dekker, New York.

Kaplan, N.L. and J.F.Y. Brookfield. 1983. Transposable elements in Mendelian populations. III. Statistical results. Genetics 104:485–495.

Kaplan, N., T. Darden and C.H. Langley. 1985. Evolution and extinction of transposable elements in Mendelian populations. Genetics 109:459–480.

Kaplan, N.L. and R.R. Hudson. 1987. On the divergence of genes in multigene families. Theor. Pop. Biol. 31:178–194.

Kaplan, N. and C.H. Langley. 1979. A new estimate of sequence divergence of mitochondrial DNA using restriction endonuclease mapping. J. Mol. Evol. 13:295–304.

Kaplan, N. and K. Risko. 1981. An improved method for estimating sequence divergence of DNA using restriction endonuclease mappings. J. Mol. Evol. 17:156–162.

Karlin, S. 1969. *Equilibrium Behaviour of Population Genetics Models with Non-Random Mating*. Gordon and Breach, London.

Karlin, S. 1975. General two-locus selection models: Some objectives, results and interpretations. Theor. Pop. Biol. 7:364–398.

Karlin, S. 1979. Principles of polymorphism and epistasis for multilocus systems. Proc. Natl. Acad. Sci. USA 76:541–545.

Karlin, S. 1982. Classifications of selection-migration structures and conditions for a protected polymorphism. Evol. Biol. 14:61–204.

Karlin, S. 1986. Comparative analysis of structural relationships in DNA and protein sequences. pp. 329–363. *In* S. Karlin and E. Nevo (eds.), *Evolutionary Processes and Theory*. Academic Press, New York.

Karlin, S. and H. Avni. 1981. Analysis of central equilibria in multilocus systems: A generalized symmetric viability regime. Theor. Pop. Biol. 20:241–280.

Karlin, S. and D. Carmelli. 1975. Numerical studies on two-locus selection models with general viabilities. Theor. Pop. Biol. 7:364–398.

Karlin, S. and M.W. Feldman. 1968. Further analysis of negative assortative mating. Genetics 59:117–136.

Karlin, S. and M.W. Feldman. 1970. Linkage and selection: Two-locus symmetric viability model. Theor. Pop. Biol. 1:39–71.

Karlin, S. and M.W. Feldman. 1981. A theoretical and numerical assessment of genetic variability. Genetics 97:475–493.

Karlin, S. and S. Lessard. 1983. On the optimal sex ratio. Proc. Natl. Acad. Sci. USA 80:5931–5935.

Karlin, S. and U. Liberman. 1979. Central equilibria in multilocus systems. I. Generalized nonepistatic regimes. Genetics 91:777–798.

Karlin, S. and U. Liberman. 1982. The reduction property for central polymorphisms in nonepistatic systems. Theor. Pop. Biol. 22:69–95.

Karlin, S. and J. McGregor. 1972. Addendum to a paper of W. Ewens. Theor. Pop. Biol. 3:113–116.

Karlin, S. and J. McGregor. 1974. Towards a a theory of the evolution of modifier genes. Theor. Pop Biol. 5:59–103.

Karlin, S., M. Morris, G. Ghandour and M.-Y. Leung. 1988. Efficient algorithms for molecular sequence analysis. Proc. Natl. Acad. Sci. USA 85:841–845.

Karlin, S. and J. Raper. 1979. Sexual selection encounter models. Theor. Pop. Biol. 15:246–256.

Karlin, S. and F.M. Scudo. 1969. Assortative mating based on phenotype. II. Two autosomal alleles without dominance. Genetics 63:499–510.

Karlin, S. and H.M. Taylor. 1975. *A First Course in Stochastic Processes*, 2nd Ed. Academic Press, New York.

Karn, M.N. and L.S. Penrose. 1951. Birth weight and gestation time in relation to maternal age, parity and infant survival. Ann. Eugen. 16:147–164.

Kazazian, H.H., A. Chakravarti, S.H. Orkin and S.E. Antonarakis. 1983. DNA polymorphism in the human β-globin gene cluster. pp. 137–146. *In* M. Nei and R.K. Koehn (eds.), *Evolution of Genes and Proteins*. Sinauer Associates, Sunderland, MA.

Kazazian, H.H., C. Wong, H. Youssoufian, A.F. Scott, D.G. Phillips and S.E. Antonarakis. 1988. Haemophilia A resulting from *de novo* insertion of L1 sequences represents a novel mechanism for mutation in man. Nature 332:164–166.

Keightley, P.D. and H. Kacser. 1987. Dominance, pleiotropy and metabolic structure. Genetics 117:319–329.

Keith, T.P. 1983. Frequency distribution of *esterase-5* alleles in two populations of *Drosophila pseudoobscura*. Genetics 105:135–155.

Keith, T.P., L.D. Brooks, R.C. Lewontin, J.C. Martinez-Cruzado and D.L. Rigby. 1985. Nearly identical distributions of xanthine dehydrogenase in two populations of *Drosophila pseudoobscura*. Mol. Biol. Evol. 2:206–216.

Kempthorne, O. 1978. Logical, epistemological and statistical aspects of nature-nurture data interpretation. Biometrics 34:1–23.

Kerr, W.E. 1969. Some aspects of the evolution of social bees. Evol. Biol. 3:119–268.

Kerr, W.E. 1974. Advances in cytology and genetics of bees. Annu. Rev. Entomol. 19:253–268.

Kerr, W.E. and S. Wright. 1954. Experimental studies of the distribution of gene frequencies in very small populations of *Drosophila melanogaster*. I. *forked*. Evolution 8:172–177.

Kessler, S. 1966. Selection for and against ethological isolation between *Drosophila pseudoobscura* and *Drosophila persimilis*. Evolution 20:634–645.

Kettlewell, H.B.D. 1973. *The Evolution of Melanism: The Study of a Recurring Necessity*. Clarendon Press, Oxford.

Keyfitz, N. 1977. *Introduction to the Mathematics of Populations, with Revisions*. Addison-Wesley, Reading, MA.

Kidd, K.K. 1987. Population genetics of a disease. Nature 327:282–283.

Kidwell, J.F., M.T. Clegg, F.M. Stewart and T. Prout. 1977a. Regions of stable equilibria for models of differential selection in the two sexes under random mating. Genetics 85:171–183.

Kidwell, M.G., J.F. Kidwell and J.A. Sved. 1977b. Hybrid dysgenesis in *Drosophila melanogaster*: A syndrome of aberrant traits including mutation, sterility, and male recombination. Genetics 86:813–833.

Kim, J. and M.A. Burgman. 1988. Accuracy of phylogenetic estimation methods under unequal evolutionary rates. Evolution 42:596–602.

Kimura, M. 1954. Process leading to quasi-fixation of genes in natural populations due to random fluctuations of selection intensities. Genetics 39:280–295.

Kimura, M. 1955. Solution of a process of random genetic drift with a continuous model. Proc. Natl. Acad. Sci. USA 41:144–150.

Kimura, M. 1964. *Diffusion Models in Population Genetics*. Methuen, London.

Kimura, M. 1965. A stochastic model concerning the maintenance of genetic variability in quantitative characters. Proc. Natl. Acad. Sci. USA 54:731–736.

Kimura, M. 1968. Evolutionary rate at the molecular level. Nature 217:624–626.

Kimura, M. 1969. The number of heterozygous nucleotide sites maintained in a finite population due to steady flux of mutations. Genetics 61:893–903.

Kimura, M. 1971. Theoretical foundation of population genetics at the molecular level. Theor. Pop. Biol. 2:174–208.

Kimura, M. 1976. Population genetics and molecular evolution. Johns Hopkins Medical J. 138:253–261.

Kimura, M. 1983. *The Neutral Theory of Molecular Evolution*. Cambridge University Press, Cambridge.

Kimura, M. and J.F. Crow. 1963. The measurement of effective population number. Evolution 17:279–288.

Kimura, M. and J.F. Crow. 1978. Effect of overall phenotypic selection on genetic change at individual loci. Proc. Natl. Acad. Sci. USA 75:6168–6171.

Kimura, M. and T. Maruyama. 1971. Pattern of neutral polymorphism in a geographically structured population. Genet. Res. 18:125–131.

Kimura, M. and T. Ohta. 1971. *Theoretical Aspects of Population Genetics*. Princeton University Press, Princeton, New Jersey.

Kimura, M. and G.H. Weiss. 1964. The stepping stone model of population structure and the decrease of genetic correlation with distance. Genetics 49:561–576.

King, J.L. and T.H. Jukes. 1969. Non-Darwinian evolution. Science 164:788–798.

Kingman, J.F.C. 1961. A mathematical problem in population genetics. Proc. Camb. Phil. Soc. 57:574–582.

Kingman, J.F.C. 1978. A simple model for the balance between selection and mutation. J. Appl. Probab. 15:1–12.

Kingman, J.F.C. 1980. *Mathematics of Genetic Diversity*. Society for Industrial and Applied Mathematics, Philadelphia.

Kingman, J.F.C. 1982. On the genealogy of large populations. J. Appl. Prob. 19A:27–43.

Kirkpatrick, M. 1982. Sexual selection and the evolution of female choice. Evolution 36:1–12.

Kirkpatrick, M. 1985. Evolution of female choice and male parental investment in polygynous species: The demise of the "sexy son." Am. Nat. 125:788–810.

Kiyasu, P.K. and M.G. Kidwell. 1984. Hybrid dysgenesis in *Drosophila melanogaster*: The evolution of mixed P and M populations maintained at high temperature. Genet. Res. 44:251–259.

Klarenberg, A.J. and W. Scharloo. 1986. Nonrandom association between structural *Amy* and regulatory *map* variants in *Drosophila melanogaster*. Genetics 114:875–884.

Kleckner, N. 1977. Translocatable elements in procaryotes. Cell 11:11–23.

Klitz, W. and G. Thomson. 1987. Disequilibrium pattern analysis. II. Application to Danish HLA A and B locus data. Genetics 116:633–643.

Kluge, R. and H. Geldermann. 1982. Effects of marker chromosome sections on quantitative traits in the mouse. Theor. Appl. Genet. 62:1–4.

Knight, G.R., A. Robertson and C.H. Waddington. 1956. Selection for sexual isolation within a species. Evolution 10:14–22.

Koch, R.E. 1971. The influence of neighboring base pairs upon base-pair substitution mutation rates. Proc. Natl. Acad. Sci. USA 68:773–776.

Koehn, R.K., W.J. Diehl and T.M. Scott. 1988. The differential contribution by individual enzymes of glycolysis and protein catabolism to the relationship between heterozygosity and growth rate in the coot clam *Mulinia lateralis*. Genetics 118:121–130.

Koehn, R.K. and P.M. Gaffney. 1984. Genetic heterozygosity and growth rate in *Mytilus edulis*. Mar. Biol. 82:1–7.

Koehn, R.K. and T.J. Hilbish. 1987. The adaptive importance of genetic variation. Amer. Sci. 75:134–141.

Koehn, R.K., R.I.E. Newell and F. Immerman. 1980. Maintenance of an aminopeptidase allele frequency cline by natural selection. Proc. Natl. Acad. Sci. USA 77:5385–5389.

Kohn, L.A. and W.R. Atchley. 1988. How similar are genetic correlation structures? Data from mice and rats. Evolution 42:467–481.

Komma, D.J. and S.A. Endow. 1987. Incomplete Y chromosomes promote magnification in male and female *Drosophila*. Proc. Natl. Acad. Sci. USA 84:2382–2386.

Koop, B.F., M. Goodman, P. Xu, K. Chan and J.L. Slightom. 1986. Primate η-globin DNA sequences and man's place among the great apes. Nature 319:234–237.

Kreitman, M. 1983. Nucleotide polymorphism at the *alcohol dehydrogenase* locus of *Drosophila melanogaster*. Nature 304:412–417.

Kreitman, M. and M. Aguade. 1986a. Genetic uniformity in two populations of *Drosophila melanogaster* revealed by filter hybridization of four-nucleotide-recognizing restriction enzyme digests. Proc. Natl. Acad. Sci. USA 83:3562–3566.

Kreitman, M. and M. Aguade. 1986b. Excess polymorphism at the *Adh* locus in *Drosophila melanogaster*. Genetics 114:93–110.

Kricker, M. and B.G. Hall. 1984. Directed evolution of cellobiose utilization in *Escherichia coli* K12. Mol. Biol. Evol. 1:171–182.

Kricker, M. and B.G. Hall. 1987. Biochemical genetics of the cryptic gene system for cellobiose utilization in *Escherichia coli* K12. Genetics 115:419–429.

Kubitschek, H.E. 1974. Operation of selection pressure on microbial populations. pp. 105–130. *In*

Evolution in the Microbial World, 24th Symposium of the Society for General Microbiology.

Lake, J.A. 1987. Determining evolutionary distances from highly diverged nucleic acid sequences: Operator metrics. J. Mol. Evol. 26:59–73.

Lande, R. 1975. The maintenance of genetic variability by mutation in a polygenic character with linked loci. Genet. Res. 26:221–234.

Lande, R. 1976. Natural selection and random genetic drift in phenotypic evolution. Evolution 30:314–334.

Lande, R. 1977. The influence of the mating system on the maintenance of genetic variability in polygenic characters. Genetics 86:485–498.

Lande, R. 1979. Quantitative genetic analysis of multivariate evolution, applied to brain:body size allometry. Evolution 33:402–416.

Lande, R. 1980. The genetic covariance between characters maintained by pleiotropic mutations. Genetics 94:203–215.

Lande, R. 1981. The minimum number of genes contributing to quantitative variation between and within populations. Genetics 99:541–553.

Lande, R. 1985. Expected time for random genetic drift of a population between stable phenotypic states. Proc. Natl. Acad. Sci. USA 82:7641–7645.

Lande, R. and S.J. Arnold. 1983. The measurement of selection on correlated characters. Evolution 37:1210–1226.

Lande, R. and S.J. Arnold. 1985. Evolution of mating preferences and sexual dimorphism. J. Theor. Biol. 117:651–664.

Lander, E.S. and P. Green. 1987. Construction of multilocus genetic linkage maps in humans. Proc. Natl. Acad. Sci. USA 84:2363–2367.

Landsmann, J., E.S. Dennis, T.J.V. Higgins, C.A. Appleby, A.A. Kortt and W.J. Peacock. 1986. Common evolutionary origin of legume and nonlegume plant hemoglobins. Nature 324:166–168.

Langley, C.H. and C.F. Aquadro. 1987. Restriction-map variation in natural populations of *Drosophila melanogaster*: *White*-locus region. Mol. Biol. Evol. 4:651–663.

Langley, C.H., J.F.Y. Brookfield and N. Kaplan. 1983. Transposable elements in Mendelian populations. I. A theory. Genetics 104:457–471.

Langley, C.H. and W.M. Fitch. 1974. An examination of the constancy of the rate of molecular evolution. J. Mol. Evol. 3:161–177.

Langley, C.H., K. Ito and R.A. Voelker. 1977. Linkage disequilibrium in natural populations of *Drosophila melanogaster*. Seasonal variation. Genetics 86:447–454.

Langley, C.H., A.E. Shrimpton, T. Yamazaki, N. Miyashita, Y. Matsuo and C.F. Aquadro. 1988. Naturally-occurring variation in the restriction map of the *Amy* region of *Drosophila melanogaster*. Genetics 119:619–629.

Langley, C.H., D.B. Smith and F.M. Johnson. 1978. Analysis of linkage disequilibria between allozyme loci in natural populations of *Drosophila melanogaster*. Genet. Res. 32:215–230.

Langley, C.H., Y.N. Tobari and K.-I. Kojima. 1974. Linkage disequilibrium in natural populations of *Drosophila melanogaster*. Genetics 78:921–936.

Lansman, R.A., R.O Shade, T.A. Grigliatti and H.W. Brock. 1987. Evolution of *P* transposable elements: Sequence of *Drosophila nebulosa P* elements. Proc. Natl. Acad. Sci. USA 84:6491–6495.

LaPorte, D.C., K. Walsh and D.E. Koshland. 1984. The branch point effect: Ultrasensitivity and subsensitivity to metabolic control. J. Biol. Chem. 259:14068–14075.

Laski, F.A., D.C. Rio and G.M. Rubin. 1986. Tissue specificity of *Drosophila P* element transposition is regulated at the level of mRNA splicing. Cell 44:7–19.

Latorre, A., A. Moya and F.J. Ayala. 1986. Evolution of mitochondrial DNA in *Drosophila subobscura*. Proc. Natl. Acad. Sci. USA 83:8649–8653.

Latter, B.D.H. 1973. The island model of population differentiation: A general solution. Genetics 73: 147–157.

Laurie-Ahlberg, C.C., G. Maroni, G.C. Bewley, J.C. Lucchesi and B.S. Weir. 1980. Quantitative genetic variation of enzyme activities in natural populations of *Drosophila melanogaster*. Proc. Natl. Acad. Sci. USA 77:1073–1077.

Laurie-Ahlberg, C.C. and L.F. Stam. 1987. Use of *P*-element mediated transformation to identify the molecular basis of naturally occurring variants affecting *Adh* expression in *Drosophila melanogaster*. Genetics 115:129–140.

Laurie-Ahlberg, C.C., A.N. Wilton, J.W. Curtsinger and T.H. Emigh. 1982. Naturally occurring enzyme activity variation in *Drosophila melanogaster*. I. Sources of variation for 23 enzymes. Genetics 102:191–206.

Laurie-Ahlberg, C.C. P.T. Barnes, J.W. Curtsinger, T.H. Emigh, B. Karlin, R. Morris, R.A. Norman and A.N. Wilton. 1985. Genetic variability of flight metabolism in *Drosophila melanogaster* II. Relationship between power output and enzyme activity levels. Genetics 111:845–868.

Ledig, F.T., R.P. Guries and B.A. Bonefield. 1983. The relation of growth to heterozygosity in pitch pine. Evolution 37:1227–1238.

Lee, G.M. and J.E. Wright. 1981. Mitotic and meiotic analyses of brook trout, *Salvelinus fontinalis*. J. Hered. 72:321–327.

Lees, D.R. and E.R. Creed. 1977. The genetics of the insularia forms of the peppered moth, *Biston betularia*. Heredity 39:67–73.

Leigh, E.G. Jr. 1977. How does selection reconcile individual advantage with the good of the group? Proc. Natl. Acad. Sci. USA 74:4542–4546.

Leigh Brown, A.J. and C.H. Langley. 1979. Reevaluation of level of genic heterozygosity in natural populations of *Drosophila melanogaster* by two-dimensional electrophoresis. Proc. Natl. Acad. Sci. USA 76:2381–2384.

Lenski, R.E. 1984. Two-step resistance by *Escherichia coli* B to bacteriophage T2. Genetics 107:1–7.

Lenski, R.E. 1987. Dynamics of interactions between bacteria and virulent bacteriophage. Adv. Microb. Ecol. 10:1–44.

Lenski, R.E. 1988. Experimental studies of pleiotropy and epistasis in *Escherichia coli*. I. Variation in competitive fitness among mutants resistant to virus T4. Evolution 42:425–432.

Lenski, R.E. and B.R. Levin. 1985. Constraints on the coevolution of bacteria and virulent phage: A model, some experiments and predictions for natural communities. Am. Nat. 125:585–602.

Lerner, I.M. 1958. *The Genetic Basis of Selection*. John Wiley and Sons, New York.

Levene, H. 1953. Genetic equilibrium when more than one niche is available. Am. Nat. 87:331–333.

Levin, B.R. and W.L. Kilmer. 1975. Interdemic selection and the evolution of altruism: A computer simulation study. Evolution 28:527–545.

Levin, B.R. and R.E. Lenski. 1983. Coevolution in bacteria and their viruses and plasmids. pp. 99–127. *In* D.J. Futuyma and M. Slatkin (eds.), *Coevolution*. Sinauer Associates, Sunderland, MA.

Levin, B.R., F.M. Stewart and L. Chao. 1977. Resource limited growth, competition and predation: A model and some experimental studies with bacteria and bacteriophage. Am. Nat. 111:3–24.

Levin, D.A. 1978. Genetic variation in annual *Phlox*: Self-compatible versus self-incompatible species. Evolution 32:245–263.

Levin, D.A. and H.W. Kerster. 1974. Gene flow in seed plants. Evol. Biol. 7:139–220.

Levings, C.S. III. 1983. The plant mitochondrial genome and its mutants. Cell 32:659–661.

Levitt, P.R. 1975. General kin selection models for genetic evolution of sib altruism in diploid and haplodiploid species. Proc. Natl. Acad. Sci. USA 72:4531–4535.

Levy, M. 1976. Altered glycoflavone expression in induced autotetraploids of *Phlox drummondii*. Biochem. Syst. and Ecol. 4:249–254.

Lewin, B. 1987. *Genes III*. Wiley, New York.

Lewontin, R.C. 1955. The effects of population density and composition on viability in *Drosophila melanogaster*. Evolution 9:27–41.

Lewontin, R.C. 1958. A general method for investigating the equilibrium of gene frequency in a population. Genetics 43:419–434.

Lewontin, R.C. 1964. The interaction of selection and linkage. I. General considerations; heterotic models. Genetics 49:49–67.

Lewontin, R.C. 1965. Selection in and of populations. pp. 292–311. *In* J.A. Moore (ed.), *Ideas in Modern Biology*. Natural History Press, New York.

Lewontin, R.C. 1968. The effect of differential viability on the population dynamics of *t* alleles in the house mouse. Evolution 22:262–273.

Lewontin, R.C. 1972. The apportionment of human diversity. Evol. Biol. 6:381–398.

Lewontin, R.C. 1974a. *The Genetic Basis of Evolutionary Change*. Columbia University Press, New York.

Lewontin, R.C. 1974b. The analysis of variance and the analysis of causes. Am. J. Hum. Gen. 26:400–411.

Lewontin, R.C. 1985. Population genetics. Annu. Rev. Genet. 19:81–102.

Lewontin, R.C. and C.C. Cockerham. 1959. The goodness-of-fit test for detecting natural selection in random mating populations. Evolution 13:561–564.

Lewontin, R.C. and L.C. Dunn. 1960. The evolutionary dynamics of a polymorphism in the house mouse. Genetics 45:705–722.

Lewontin, R.C., L.R. Ginzburg and S.D. Tuljapurkar. 1978. Heterosis as an explanation for large amounts of genic polymorphism. Genetics 88:149–170.

Lewontin, R.C. and J.L Hubby. 1966. A molecular approach to the study of genic heterozygosity in natural populations II. Amount of variation and degree of heterozygosity in natural populations of *Drosophila pseudoobscura*. Genetics 54:595–609.

Lewontin, R.C. and K.-I. Kojima. 1960. The evolutionary dynamics of complex polymorphisms. Evolution 14:458–472.

Lewontin, R.C. and M.J.D. White. 1960. Interaction between inversion polymorphisms of two chromosome pairs in the grasshopper *Moraba scurra*. Evolution 14:116–129.

Li, C.C. 1967a. Castle's early work on selection and equilibrium. Am. J. Hum. Genet. 19:70–74.

Li, C.C. 1967b. Fundamental theorem of natural selection. Nature 214:505–506.

Li, C.C. 1969. Population subdivision with respect to multiple alleles. Ann. Hum. Genet. 33:23–29.

Li, W.-H. 1986. Evolutionary change of restriction cleavage sites and phylogenetic inference. Genetics 113:187–213.

Li, W.-H. and T. Gojobori. 1983. Rapid evolution of goat and sheep globin genes following gene duplication. Mol. Biol. Evol. 1:94–108.

Li, W.-H., T. Gojobori and M. Nei. 1981. Pseudogenes as a paradigm of neutral evolution. Nature 292:237–239.

Li, W.-H., C.-C. Luo and C.-I. Wu. 1985a. Evolution of DNA sequences. pp. 1–94. *In* R.J. MacIntyre (ed.), *Molecular Evolutionary Genetics*. Plenum, New York.

Li, W.-H., M. Tanimura and P.M. Sharp. 1988. Rates and dates of divergence between AIDS virus nucleotide sequences. Mol. Biol. Evol. 5:313–330.

Li, W.-H. and C.-I. Wu. 1987. Rates of nucleotide substitution are evidently higher in rodents than in man. Mol. Biol. Evol. 4:74–77.

Li, W.-H., C.-I. Wu and C.-C. Luo. 1985b. Nonrandomness of point mutations as reflected in nucleotide substitutions in pseudogenes and its evolutionary implications. J. Mol. Evol. 21:58–71.

Liberman, U. and M.W. Feldman. 1985. Modifiers of mutation rate: A general reduction principle. Theor. Pop. Biol. 30:125–142.

Lofsvold, D. 1986. Quantitative genetics of morphological differentiation in *Peromyscus*. I. Tests of the homogeneity of genetic covariance structure among species and subspecies. Evolution 40:559–573.

Lofsvold, D. 1988. Quantitative genetics of morphological differentiation in *Peromyscus*. II. Analysis of selection and drift. Evolution 42:54–67.

Long, J.C. 1986. The allelic correlation structure of Gainj and Kalam speaking people. I. The estimation and interpretation of Wright's *F* statistics. Genetics 112:629–647.

Long, J.C., N.M. Naidu, H.W. Mohrenweiser, H. Gershowitz, P.L. Johnson, J.W. Wood and P.E. Smouse. 1986. Genetic characterization of Gainj and Kalam speaking people of Papua New Guinea. Am. J. Phys. Anthropol. 70:75–96.

Loukas, M., C.B. Krimbas and Y. Vergini. 1979. The genetics of *Drosophila subobscura* populations. IX. Studies of linkage disequilibrium in four natural populations. Genetics 93:497–523.

Luckinbill, L.S., R. Arking, M.J. Clare, W.C. Cirocco and S.A. Buck. 1984. Selection for delayed senescence in *Drosophila melanogaster*. Evolution 38:996–1003.

Luckinbill, L.S. and M.J. Clare. 1985. Selection for life span in *Drosophila melanogaster*. Heredity 55:9–18.

Lynch, M. 1985. Spontaneous mutations for life-history characters in an obligate parthenogen. Evolution 39:804–818.

Lynch, C.B. and D.S. Sulzbach. 1984. Quantitative genetic analysis of temperature regulation in *Mus musculus*. II. Diallele analysis of individual traits. Evolution 38:527–540.

Lynch, M. 1987. The consequences of fluctuating selection for isozyme polymorphisms in *Daphnia*. Genetics 115:657–669.

Lynch, M. 1988a. The divergence of neutral quantitative characters among partially isolated populations. Evolution 42:455–466.

Lynch, M. 1988b. The rate of polygenic mutation. Genet. Res. Camb. 51:137–148.

Lynch, M. and W.G. Hill. 1986. Phenotypic evolution by neutral mutation. Evolution 40:915–935.

Lyttle, T.W. 1979. Experimental population genetics of meiotic drive systems. II. Accumulation of genetic modifiers of *Segregation Distorter* (*SD*) in laboratory populations. Genetics 91:339–35.

MacCluer, J.W., A.J. Boyce, B. Dyke, L.R. Weitkamp, D.W. Pfenning and C.J. Parsons. 1983. Inbreeding and pedigree structure in Standardbred horses. J. Hered. 74:394–399.

Mackay, T.F.C. 1984. Jumping genes meet abdominal bristles: Hybrid dysgenesis-induced quantitative variation in *Drosophila melanogaster*. Genet. Res. 44:231–237.

Mackay, T.F.C. 1985. Transposable element induced response to artificial selection in *Drosophila melanogaster*. Genetics 111:351–374.

MacNair, M. 1979. The genetics of copper tolerance in the yellow monkey flower, *Mimulus guttatus*. I. Crosses to nontolerants. Genetics 91:553–563.

MacRae, A.F. and W.W. Anderson. 1988. Evidence for non-neutrality of mitochondrial DNA haplotypes in *Drosophila pseudoobscura*. Genetics 120:485–494.

Maeda, N. and O. Smithies. 1986. The evolution of multigene families: Human haptoglobin genes. Annu. Rev. Genet. 20:81–108.

Mandel, S.P.H. 1959. Stable equilibrium at a sex-linked locus. Nature 183:1347–1348.

Mandel, S.P.H. 1971. Owen's model of a genetical system with differential viability between sexes. Heredity 26:49–63.

Mandel, S.P.H. 1980. The triangle inequality in balanced genetical polymorphisms. Genetics 96:557–559.

Manly, B.F.J. 1985. *The Statistics of Natural Selection on Animal Populations*. Chapman and Hall, London.

Margot, J.B., G.W. Demers and R.C. Hardison. In press. Complete nucleotide sequence of the rabbit β-like globin gene cluster: Analysis of intergenic sequences and comparison with the human β-like globin gene cluster. Mol. Biol. Evol.

Marinkovic, D., F.J. Ayala and M. Andjelkovic. 1978. Genetic polymorphism and phylogeny of *Drosophila subobscura*. Evolution 32:164–173.

Markow, T.A. 1978. A test for the rare male mating advantage in coisogenic strains of *Drosophila melanogaster*. Genet. Res. 32:123–127.

Markow, T.A., R.C. Richmond, L. Mueller, I. Sheer, S. Roman, C. Laetz and L. Lorenz. 1980. Testing for rare male mating advantage among various *Drosophila melanogaster* genotypes. Genet. Res. 35:59–64.

Marks, J., J.-P. Shaw and C.-K.J. Shen. 1986. Sequence organization and genomic complexity of a primate θl gene, a novel α-globin-like gene. Nature 321:785–788.

Marks, R.W. 1982. Genetic variability for density sensitivity of three components of fitness in *Drosophila melanogaster*. Genetics 101:301–316.

Maruyama, T. 1969. Genetic correlation in the stepping stone model with non-symmetrical migration rates. J. Appl. Prob. 6:463–477.

Maruyama, T. 1970. Effective number of alleles in a subdivided population. Theor. Pop. Biol. 1:273–306.

Maruyama, T. 1971. An invariant property of a structured population. Genet. Res. 18:81–84.

Maruyama, T. 1972a. Rate of decrease of genetic variability in a two-dimensional continuous population of finite size. Genetics 70:639–651.

Maruyama, T. 1972b. Some invariant properties of a geographically structured finite population: distribution of heterozygotes under irreversible mutation. Genet. Res. 20:141–149.

Maruyama, T. 1972c. Distribution of gene frequencies in a geographically structured finite population. I. Distribution of neutral genes of of genes with small effect. Ann. Hum. Genet. 35:411–423.

Maruyama, T. and M. Kimura. 1980. Genetic variability and effective population size when local extinction and recolonization of subpopulations are frequent. Proc. Natl. Acad. Sci. USA 77:6710–6714.

Matessi, C. and S.D. Jayakar. 1976. Conditions for the evolution of altruism under Darwinian selection. Theor. Pop. Biol. 9:360–387.

Matessi, C. and S. Karlin. 1986. Altruistic behavior in sibling groups with unrelated intruders. pp. 689–724. *In* S. Karlin and E. Nevo (eds.), *Evolutionary Processes and Theory*. Academic Press, New York.

Maxam, A.M. and W. Gilbert. 1977. A new method for sequencing DNA. Proc. Natl. Acad. Sci. USA 74:560–564.

Maxson, L.R. 1984. Molecular probes of phylogeny and biogeography in toads of the widespread genus *Bufo*. Mol. Biol. Evol. 1:345–356.

Maxson, R.D. and L.R. Maxson. 1986. Micro-complement fixation: A quantitative estimator of protein evolution. Mol. Biol. Evol. 3:375–388.

Maynard Smith, J. 1962. Disruptive selection, polymorphism, and sympatric speciation. Nature 195:60–62.

Maynard Smith, J. 1966. Sympatric speciation. Am. Nat. 100:637–650.

Maynard Smith, J. 1982. *Evolution and the Theory of Games*. Cambridge University Press, Cambridge.

Maynard Smith, J. and K.C. Sondhi. 1961. The genetics of a pattern. Genetics 45:1039–1050.

Mayr, E. 1963. *Animal Species and Evolution*. Harvard University Press, Cambridge, MA.

Mayr, E. 1970. *Populations, Species and Evolution*. Harvard University Press, Cambridge, MA.

Mazer, S.J. 1987. The quantitative genetics of life history and fitness components in *Raphanus raphanistrum* L. (Brassicaceae): Ecological and evolutionary consequences of seed-weight variation. Am. Nat 130:891–914.

McAllister, M.K. and B.D. Roitberg. 1987. Adaptive suicidal behavior in pea aphids. Nature 328:797–799.

McCarry, J.R. and K. Thomas. 1987. Human testis-specific PGK gene lacks introns and possesses characteristics of a processed gene. Nature 326:501–505.

McCauley, D.E. and M.J. Wade. 1980. Group selection: The genetic and demographic basis for the phenotypic differentiation of small populations of *Tribolium castaneum*. Evolution 34:813–820.

McClintock, B. and E.F. Keller. 1983. *A Feeling for the Organism: The Life and Work of Barbara McClintock*. W.H. Freeman, San Francisco.

McConnel, T.J., W.S. Talbot, R.A. McIndoe and E.K. Wakeland. 1988. The origin of MHC class II gene polymorphism within the genus *Mus*. Nature 332:651–653.

McLellan, T. and S. Inouye. 1986. The sensitivity of isoelectric focusing and electrophoresis in the detection of sequence differences in proteins. Biochem. Genet. 24:571–577.

McPheron, B.A., D. Courtney Smith and S.H. Berlocher. 1988. Microgeographic genetic variation in

the apple maggot *Rhagoletis pomonella*. Genetics 119:445–451.

Medawar, P.B. 1952. *An Unsolved Problem in Biology*. H.K. Lewis, London.

Meynell, G.G. 1972. *Bacterial Plasmids*. M.I.T. Press, Cambridge, MA.

Michaelis, L. and M.L. Menten. 1913. Die Kinetik der Invertinwirkung. Biochem. Z. 49:333–369.

Michod, R.E. 1982. The theory of kin selection. Annu. Rev. Ecol. Syst. 13:23–55.

Michod, R.E. and B.R. Levin. 1987. *The Evolution of Sex: An Examination of Current Ideas*. Sinauer Associates, Sunderland, MA.

Miklos, G.L.G. 1985. Localized highly repetitive DNA sequences in vertebrate and invertebrate genomes. pp. 241–321. *In* R.J. MacIntyre (ed.), *Molecular Evolutionary Genetics*. Plenum, New York.

Milkman, R. 1973. Electrophoretic variation in *Escherichia coli* from natural sources. Science 182:1024–1026.

Miller, R.H. 1988. Human immunodeficiency virus may encode a novel protein on the genomic DNA plus strand. Science 239:1420–1422.

Mitchell-Olds, T. 1986. Quantitative genetics of survival and growth in *Impatiens capensis*. Evolution 40:107–116.

Mitchell-Olds, T. and J.J. Rutledge. 1986. Quantitative genetics in natural populations: A review of the theory. Am. Nat. 127:379–402.

Mitchell-Olds, T. and R.G. Shaw. 1987. Regression analysis of natural selection: Statistical inference and biological interpretation. Evolution 41:1149–1161.

Mitton, J.B., C. Carey and T.D. Kocher. 1986. The relation of enzyme heterozygosity to standard and active oxygen consumption and body size of tiger salamanders, *Ambystoma tigrinum*. Physiol. Zool. 59:574–582.

Mitton, J.B. and M.C. Grant. 1984. Associations among protein heterozygosity, growth rate and developmental homeostasis. Annu. Rev. Ecol. Syst. 15:479–499.

Møller, A.P. 1988. Female choice selects for male sexual tail ornaments in the monogamous swallow. Nature 332:640–644.

Monnerot, M., J.-C. Mounolou and M. Solignac. 1984. Intra-individual length heterogeneity of *Rana esculenta* mitochondrial DNA. Biol. Cell 52:213–218.

Montgomery, E.A., B. Charlesworth and C.H. Langley. 1987. A test for the role of natural selection in the stabilization of transposable element copy number in a population of *Drosophila melanogaster*. Genet. Res. 49:31–41.

Montgomery, E.A. and C.H. Langley. 1983. Transposable elements in Mendelian populations. II. Distribution of three *copia*-like elements in a natural population of *Drosophila melanogaster*. Genetics 104:473–483.

Moran, P.A.P. 1962. *The Statistical Processes of Evolutionary Theory*. Clarendon Press, Oxford.

Moran, P.A.P. 1964. On the nonexistence of adaptive topographies. Ann. Hum. Genet. London 27:383–393.

Morse, B., P.G. Rothberg, V.J. South, J.M. Spandorfer and S.M. Astrin. 1988. Insertional mutagenesis of the *myc* locus in a human breast carcinoma. Nature 333:87–90.

Morton, N.E. 1961. Morbidity of children from consanguineous marriages. pp. 261–291. *In* A.G. Steinberg (ed.), *Progress in Medical Genetics*, Vol. 1. Grune and Stratton, New York.

Morton, N.E. 1978. Effect of inbreeding on IQ and mental retardation. Proc. Natl. Acad. Sci. USA 75:3906–3908.

Morton, N.E., J.F. Crow and H.J. Muller. 1956. An estimate of mutational damage in man from data on consanguineous marriages. Proc. Natl. Acad. Sci. USA 42:855–863.

Mourant, A.E., A.C. Kopec and K. Domaniewska-Sobczak. 1976. *The Distribution of the Human Blood Groups and Other Polymorphisms*, 2nd Ed. Oxford Univeristy Press, New York.

Mueller, L.D. 1986. Density-dependent rates of population growth: estimation in laboratory populations. Am. Nat. 128:282–293.

Mueller, L.D. 1987. Evolution of accelerated senescence in laboratory populations of *Drosophila*. Proc. Natl. Acad. Sci. USA 84:1974–1977.

Mueller, L.D. and F.J. Ayala. 1981. Trade-off between *r*-selection and *K*-selection in *Drosophila* populations. Proc. Natl. Acad. Sci. USA 78:1303–1305.

Mueller, L.D., L.G. Barr and F.J. Ayala. 1985a. Natural selection vs. random drift: Evidence from temporal variation in allele frequencies in nature. Genetics 111:517–554.

Mueller, L.D., B.A. Wilcox, P.R. Ehrlich, D.G. Heckel and D.D. Murphy. 1985b. A direct assessment of the role of genetic drift in determining allele frequency variation in populations of *Euphydryas editha*. Genetics 110:495–511.

Mukai, T. 1964. The genetic structure of natural populations of *Drosophila melanogaster*. I. Spontaneous mutation rate of polygenes controlling viability. Genetics 50:1–19.

Mukai, T., L.E. Mettler and S. Chigusa. 1971. Linkage disequilibrium in a local population of *Drosophila melanogaster*. Proc. Natl. Acad. Sci. USA 68:1065–1069.

Mukai, T., T.K. Watanabe and O. Yamaguchi. 1974. The genetic structure of natural populations of *Drosophila melanogaster*. XII. Linkage disequilibrium in a large local population. Genetics 77:771–793.

Mukai, T. and O. Yamaguchi. 1974. The genetic structure of natural populations of *Drosophila melanogaster*. XI. Genetic variability of local populations. Genetics 76:339–366.

Mukai, T. and T. Yamazaki. 1980. Test for selection on polymorphic isozyme genes using the population cage method. Genetics 96:537–542.

Mulcahy, D.L. and S.M. Kaplan. 1979. Mendelian ratios despite nonrandom fertilization? Am. Nat. 113:419–425.

Muller, H.J. 1927. Artificial transmutation of the gene. Science 66:84–87.

Nagylaki, T. 1984. The evolution of multigene families under intrachromosomal gene conversion. Genetics 106:529–548.

Nagylaki, T. 1987. Evolution under fertility and viability selection. Genetics 115:367–375.

Nakajima, K., U. Desselberger and P. Palease. 1978. Recent human influenza A (H1N1) viruses are closely related genetically to strains isolated in 1950. Nature 274:334–339.

Nakamura, Y., M. Loppert, P. O'Connell, R. Wolff, T. Holm, M. Culver, C. Martin, E. Fujimoto, M. Hoff, E. Kumlin and R. White. 1987. Variable numbers of tandem repeat (VNTR) markers for human gene mapping. Science 235:1616–1622.

Nasrallah, J.B., T.-H. Kao, M.L. Goldberg and M.E. Nasrallah. 1985. A cDNA clone encoding an *S*-locus-specific glycoprotein from *Brassica oleracea*. Nature 318:263–267.

Nathans, J., T.P. Piantanida, R.L. Eddy, T.B. Shows and D.S. Hogness. 1986. Molecular genetics of inherited variation in human color vision. Science 232:203–210.

Neal, N.P. 1935. The decrease in yielding capacity in advanced generations of hybrid corn. J. Am. Soc. Agron. 27:666–670.

Needleman, S.B. and C.D. Wunsch. 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins. J. Mol. Biol. 48:443–453.

Neel, J.V., C. Satoh, K. Goriki, J.-T. Asakawa, M. Fujita, N. Takahashi, T. Kageoka and R. Hazama. 1988. Search for mutations altering protein charge and/or function in children of atomic bomb survivors: Final report. Am. J. Hum. Gen. 42:663–676.

Neel, J.V., C. Satoh, K. Goriki, M. Fujita, N. Takahashi, J.-T. Asakawa and R. Hazama. 1986. The rate with which spontaneous mutation alters the electrophoretic mobility of polypeptides. Proc. Natl. Acad. Sci. USA 83:389–393.

Neel, J.V. and W.J. Schull. 1962. The effect of inbreeding on mortality and morbidity in two Japanese cities. Proc. Natl. Acad. Sci. USA 48:573–582.

Neel, J.V. and E.A. Thompson. 1978. Founder effect and the number of private polymorphisms observed in Amerindian tribes. Proc. Natl. Acad. Sci. USA 75:1904–1908.

Nei, M. 1965. Variation and covariation of gene frequencies in subdivided populations. Evolution 19:256–258.

Nei, M. 1967. Modification of linkage intensity by natural selection. Genetics 57:625–641.

Nei, M. 1973. Analysis of gene diversity in subdivided populations. Proc. Natl. Acad. Sci. USA 70:3321–3323.

Nei, M. 1975. *Molecular Population Genetics and Evolution*. American Elsevier, New York.

Nei, M. 1977. *F*-statistics and analysis of gene diversity in subdivided populations. Ann. Hum. Genet. 41:225–233.

Nei, M. 1978. The theory of genetic distance and evolution of human races. Jap. J. Human Genet. 23:341–369.

Nei, M. 1985. Human evolution at the molecular level. pp. 41–64. *In* T. Ohta and K. Aoki (eds.), *Population Genetics and Molecular Evolution*. Japanese Scientific Societies Press, Tokyo.

Nei, M. 1986. Definition and estimation of fixation indices. Evolution 40:643–645.

Nei, M. 1987. *Molecular Evolutionary Genetics*. Columbia University Press, New York.

Nei, M. and R.K. Chesser. 1983. Estimation of fixation indices and gene diversities. Ann. Hum. Genet. 47:253–259.

Nei, M., P.A. Fuerst and R. Chakraborty. 1976. Testing the neutral mutation hypothesis by distribution of single locus heterozygosity. Nature 262:491–493.

Nei, M. and W.-H. Li. 1973. Linkage disequilibrium in subdivided populations. Genetics 75:213–219.

Nei, M. and W.-H. Li. 1979. Mathematical models for studying genetic variation in terms of restriction endonucleases. Proc. Natl. Acad. Sci. USA 76:5269–5273.

Nei, M., T. Maruyama and R. Chakraborty. 1975. The bottleneck effect and genetic variability in populations. Evolution 29:1–10.

Nei, M. and A.K. Roychoudhury. 1982. Genetic relationship and evolution of human races. Evol. Biol. 14:1–59.

Nei, M. and F. Tajima. 1983. Maximum likelihood estimation of the number of nucleotide substitutions from restriction sites data. Genetics 105:207–217.

Nei, M. and F. Tajima. 1985. Evolutionary change of restriction cleavage sites and phylogenetic inference for man and apes. Mol. Biol. Evol. 2:189–205.

Nevo, E. 1978. Genetic variation in natural populations: Patterns and theory. Theor. Pop. Biol. 13:121–177.

Nevo, E. and H. Cleve. 1978. Genetic differentiation during speciation. Nature 275:125.

Newcombe, H.B. 1964. Epidemiologic studies (population, racial, family, seasonal, genetic). pp. 345–349. *In* M. Fishbein (ed.), *Papers and Discussions of the Second International Conference on Congenital Malformations*. International Medical Congress, New York.

Nichols, B.P. and C. Yanofsky. 1979. Nucleotide sequences of trpA of *Salmonella typhimurium* and *Escherichia coli*: An evolutionary comparison. Proc. Natl. Acad. Sci. USA 76:5244–5248.

Nilsson-Ehle, H. 1909. Kreuzungsuntersuchungen an Hafer und Weizen. Lunds. Univ. Aarskr. NF 5,2:1–22.

Nordskog, A.W. and F.G. Giesbrecht. 1964. Regression in egg production in the domestic fowl when selection is relaxed. Genetics 50:407–416.

Norton, H.T.J. 1928. Natural selection and Mendelian variation. Proc. Lond. Math. Soc. 28:1–45.

Nöthel, H. 1987. Adaptation of *Drosophila melanogaster* populations to high mutation pressure: Evolutionary adjustments of mutation rates. Proc. Natl. Acad. Sci. USA 84:1045–1049.

Nunney, L. 1983. Sex differences in larval competition in *Drosophila melanogaster*: The testing of a competition model and its relevance to frequency-dependent selection. Am. Nat. 121:67–93.

Nur, U., J.H. Werren, D.G. Eickbush, W.D. Burke and T.H. Eickbush. 1988. A "selfish" B chromosome that enhances its transmission by eliminating the paternal genome. Science 240:512–514.

Oakeshott, J.G. 1976. Selection at the *Adh* locus in *Drosophila melanogaster* imposed by environmental ethanol. Genet. Res. 26:265–274.

O'Brien, S.J., M.E. Roelke, L. Marker, A. Newman, C.A. Winkler, D. Meltzer, L. Colly, J.F. Evermann, M. Bush and D.E. Wildt. 1985. Genetic basis for species vulnerability in the cheetah. Science 227:1428–1434.

O'Brien, S.J., D.E. Wildt, D. Goldman, C.R. Merril and M. Bush. 1983. The cheetah is depauperate in genetic variation. Science 221:459–461.

O'Brien, S.J., D.E. Wildt, M. Bush, T.M. Caro, C. FitzGibbon, I. Aggundey and R.E. Leakey. 1987. East African cheetahs: Evidence for two population bottlenecks. Proc. Natl. Acad. Sci. USA 84:508–511.

O'Donald, P. 1978. Theoretical aspects of sexual selection. Theor. Pop. Biol. 12:298–334.

O'Donald, P. 1979. Theoretical aspects of sexual selection: Variation in threshold of female mating response. Theor. Pop. Biol. 15:191–204.

O'Donald, P. 1980. *Genetic Models of Sexual Selection*. Cambridge University Press, London.

O'Donald, P. 1987. Polymorphism and sexual selection in arctic skua. pp. 433–450. *In* F. Cooke and P.A. Buckley (eds.), *Avian Genetics*. Academic Press, New York.

Ohnishi, O. 1977. Spontaneous and ethyl methanesulfonate induced mutations controlling viability in *Drosophila melanogaster*. I. Recessive lethal mutations. Genetics 87:519–527.

Ohta, T. 1973. Slightly deleterious mutant substitutions in evolution. Nature 246:96–98.

Ohta, T. 1980. *Evolution and Variation of Multigene Families*. Springer-Verlag, Berlin.

Ohta, T. 1982. Allelic and nonallelic homology of a supergene family. Proc. Natl. Acad. Sci. USA 79:3251–3254.

Ohta, T. 1983. Theoretical study on the accumulation of selfish DNA. Genet. Res. 41:1–15.

Ohta, T. 1987. Simulating evolution by gene duplication. Genetics 115:207–213.

Ohta, T. 1988. Further simulation studies on evolution by gene duplication. Evolution 42:375–386.

Ohta, T. and M. Kimura. 1971. On the constancy of the evolutionary rate of cistrons. J. Mol. Evol. 1:18–25.

Ohta, T. and M. Kimura. 1973. A model of mutation appropriate to estimate the number of electrophoretically detectable alleles in a finite population. Genet. Res. 22:201–204.

Orgel, L.E. and F.H.C. Crick. 1980. Selfish DNA: The ultimate parasite. Nature 284:606–607.

Orkin, S.H. and H.H. Kazazian Jr. 1984. The mutation and polymorphism of the human β-globin gene and its surrounding DNA. Annu. Rev. Genet. 18:131–171.

Oster, G., I. Eshel and D. Cohen. 1977. Worker–queen conflict and the evolution of social insects. Theor. Pop. Biol. 12:49–85.

Østergaard, H. and F.B. Christiansen. 1981. Selection component analysis of natural polymorphisms using population samples including mother–offspring combinations, II. Theor. Pop. Biol. 19:378–419.

Palmer, J.D. 1985. Evolution of chloroplast and mitochondrial DNA in plants and algae. pp. 131–240. *In* R.J. MacIntyre (ed.), *Molecular Evolutionary Genetics*. Plenum, New York.

Pantazidis, A.C. and E. Zouros. 1988. Location of an autosomal factor causing sterility in *Drosophila mojavensis* males carrying the *Drosophila arizonensis* Y chromosome. Heredity 60:299–304.

Palmer, J.D. 1988. Intraspecific variation and multicirularity in *Brassica* mitochondrial DNAs. Genetics 118:341–351.

Peck, J.R. and M.W. Feldman. 1986. The evolution of helping behavior in large, randomly mixed populations. Am. Nat. 127:209–221.

Peetz, E.W., G. Thomson and P.W. Hedrick. 1986. Charge changes in protein evolution. Mol. Biol. Evol. 3:84–94.

Penrose, L.S. 1949. The meaning of fitness in human populations. Ann. Eugen. 14:301–304.

Perutz, M.F. 1983. Species adaptation in a protein molecule. Mol. Biol. Evol. 1:1–28.

Petit, C. and L. Ehrman. 1969. Sexual selection in *Drosophila*. Evol. Biol. 3:177–223.

Pichersky, E., R. Bernatzky, S.D. Tanksley and A.R. Cashmore. 1986. Evidence for selection as a mechanism in the concerted evolution of *Lycopersicon esculentum* (tomato) genes encoding the small subunit of ribulose-1,5-bisphosphate carboxylase/oxygenase. Proc. Natl. Acad. Sci. USA 83:3880–3884.

Pirchner, F. 1969. *Population Genetics in Animal Breeding*. W.H. Freeman, San Francisco.

Platt, A.P. and L.P. Brower. 1968. Mimetic versus disruptive coloration in intergrading populations of *Limenitis arthemia* and *Astyanax* butterflies. Evolution 22:699–718.

Pollak, E. 1978. With selection for fecundity the mean fitness does not necessarily increase. Genetics 90:383–389.

Pollak, E. and O. Kempthorne. 1970. Malthusian parameters in genetic populations. I. Haploid and selfing models. Theor. Pop. Biol. 1:315–345.

Pollak, E. and O. Kempthorne. 1971. Malthusian parameters in genetic populations. II. Random mating populations in infinite habitats. Theor. Pop. Biol. 2:351–390.

Powell, J.R. 1978. The founder-flush speciation theory: An experimental approach. Evolution 32:465–474.

Powell, J.R. and G.D. Amato. 1984. Population genetics of *Drosophila* amylase. V. Genetic background and selection on different carbohydrates. Genetics 106:625–629.

Powell, J.R. and M. Andjelkovic. 1983. Population genetics of *Drosophila* amylase. IV. Selection in laboratory populations maintained on different carbohydrates. Genetics 103:675–689.

Powell, J.R., A. Caccone, G.D. Amato and C. Yoon. 1986. Rates of nucleotide substitution in *Drosophila* mitochondrial DNA and nuclear DNA are similar. Proc. Natl. Acad. Sci. USA 83:9090–9093.

Powers, D.A., L. DiMichele and A.R. Place. 1983. The use of enzyme kinetics to predict differences in cellular metabolism, developmental rate and swimming performance between LDH-B genotypes of the fish *Fundulus heteroclitus*. Isozymes. Curr. Top. Biol. Med. Res. 10:147–170.

Prakash, S. 1977. Further studies on gene polymorphism in the mainbody and geographically isolated populations of *Drosophila pseudoobscura*. Genetics 85:713–719.

Prakash, S. and R.C. Lewontin. 1968. A molecular approach to the study of genic heterozygosity III. Direct evidence of coadaptation in gene arrangements of *Drosophila*. Proc. Natl. Acad. Sci. USA 59:398–405.

Press, W.H., B.P. Flannery, S.A. Teukolsky and W.T. Vetterling. 1986. *Numerical Recipes: The Art of Scientific Computing*. Cambridge University Press, Cambridge.

Price, T.D. and P.R. Grant. 1985. The evolution of ontogeny in Darwin's finches: A quantitative genetic approach. Amer. Nat. 125:169–188.

Prout, T. 1965. The estimation of fitness from genotypic frequencies. Evolution 19:546–551.

Prout, T. 1968. Sufficient conditions for multiple niche polymorphism. Am. Nat. 102:493–496.

Prout, T. 1969. The estimation of fitness from population data. Genetics 63:949–967.

Prout, T. 1971a. The relation between fitness components and population prediction in *Drosophila*. I. The estimation of fitness components. Genetics 68:127–149.

Prout, T. 1971b. The relation between fitness components and population prediction in *Drosophila*. II. Population prediction. Genetics 68:151–167.

Prout, T. and J. Bundgaard. 1977. Population genetics of sperm displacement. Genetics 85:95–124.

Prout, T., J. Bundgaard and S. Bryant. 1973. Population genetics of modifiers of meiotic drive. I. The solution of a special case and some general implications. Theor. Pop. Biol. 4:446–465.

Prout, T. and F. McChesney. 1985. Competition among immatures affects their fertility when they grow up: population dynamics. Am. Nat. 126: 521–558.

Race, R.R. and R. Sanger. 1975. *Blood Groups in Man*, 6th Ed. J.B. Lippincott, Philadelphia.

Ramshaw, J.A.M., J.A. Coyne and R.C. Lewontin. 1979. The sensitivity of gel electrophoresis as a detector of genetic variation. Genetics 93:1019–1037.

Rand, D.M. and R.G. Harrison. 1986. Mitochondrial DNA transmission genetics in crickets. Genetics 114:955–970.

Rapoport, A. and A.M. Chammah. 1965. *Prisoner's Dilemma: A Study in Conflict and Cooperation*. University of Michigan Press, Ann Arbor.

Reed, T.E. and J.V. Neel. 1959. Huntington's chorea in Michigan. Am. J. Hum. Genet. 11:107–136.

Reeder, R.H. 1984. Enhancers and ribosomal gene spacers. Cell 38:349–351.

Reznick, D.N. 1982. The impact of predation on life history evolution in Trinidadian guppies: Genetic basis of observed life history patterns. Evolution 36:1236–1250.

Reznick, D.N. and H. Bryga. 1987. Life–history evolution in guppies (*Poecilia reticulata*). 1. Phenotypic and genetic changes in an introduction experiment. Evolution 41:1370–1385.

Rice, W.R. 1986. On the instability of polygenic sex determination: The effect of sex-specific selection. Evolution 40:633–639.

Richardson, R.H. 1974. Effects of dispersal, habitat selection and competition on a speciation pattern of *Drosophila* endemic to Hawaii. pp. 140–164. *In* M.J.D. White (ed.), *Genetic Mechanisms of Speciation in Insects*. Australia and New Zealand Book Co., Sydney.

Richmond, R.C. 1972. Enzyme variability in the *Drosophila willistoni* group. III. Amounts of variability in the superspecies *D. paulistorum*. Genetics 71:87–112.

Richmond, R.C., D.G. Gilbert, K.B. Sheehan, M.H. Gromko and F.M. Butterworth. 1980. Esterase-6 and reproduction in *Drosophila melanogaster*. Science 207:1483–1485.

Rio, D.C., F.A. Laski and G.M. Rubin. 1986. Identification and immunological analysis of biologically active *Drosophila P* element transposase. Cell 44:21–32.

Riska, B., W.R. Atchley and J.J. Rutledge. 1984. A genetic analysis of targeted growth in mice. Genetics 107:79–101.

Robertson, A. 1957. Studies in quantitative inheritance. XI. Genetic and environmental correlation between body size and egg production in *Drosophila melanogaster*. J. Genet. 55:428–443.

Robertson, A. 1962. Selection for heterozygotes in small populations. Genetics 47:1291–1300.

Robinson, H.F., R.E. Comstock and P.H. Harvey. 1949. Estimates of heritability and degree of dominance in corn. Agron. J. 41:353–359.

Rohlf, F.J. and M.C. Wooten. 1988. Evaluation of the restricted maximum-likelihood method for estimating phylogenetic trees using simulated allele-frequency data. Evolution 42:581–595.

Rose, M.R. 1984. Laboratory evolution of postponed senescence in *Drosophila melanogaster*. Evolution 38:1004–1010.

Rose, M.R. and B. Charlesworth. 1981a. Genetics of life history in *Drosophila melanogaster*. I. Sib analysis of adult females. Genetics 97:173–186.

Rose, M.R. and B. Charlesworth. 1981b. Genetics of life history of *Drosophila melanogaster*. II. Exploratory selection experiments. Genetics 97:187–196.

Roughgarden, J. 1971. Density-dependent natural selection. Ecology 52:453–468.

Royden, C.S., V. Pirotta and L.Y. Jan. 1988. The *tko* locus, site of a behavioral mutation in *D. melanogaster*, codes for a protein homologous to prokaryotic ribosomal protein S12. Cell 51:165–173.

Rubin, G. 1983. Dispersed repetitive DNAs in *Drosophila*. pp. 329–361. *In* J.A. Shapiro (ed.), *Mobile Genetic Elements*. Academic Press, New York.

Rubin, G.M., M.G. Kidwell and P.M. Bingham. 1982. The molecular basis of P–M hybrid dysgenesis: The nature of induced mutations. Cell 29:987–994.

Ruddle, F.H., T.H. Roderick, T.B. Shows, P.G. Weigl, R.K. Chipman and P.K. Anderson. 1969. Measurement of genetic heterogeneity by means of enzyme polymorphism. J. Hered. 60:321–322.

Russell, W.L. 1963. The effect of radiation dose rate and fractionation on mutation in mice. pp. 205–218. *In* F.H. Sobels (ed.), *Repair from Genetic Radiation Damage*. Macmillan, New York.

Rutowski, R.L., M. Newton and J. Schaefer. 1983. Interspecific variation in the size of the nutrient investment made by male butterflies during copulation. Evolution 37:708–713.

Ruvolo, M. and T.F. Smith. 1986. Phylogeny and DNA–DNA hybridization. Mol. Biol. Evol. 3:285–289.

Ryder, L.P. and A. Svejgaard. 1981. Genetics of HLA disease association. Annu. Rev. Genet. 15:169–187.

Saag, M.S., B.H. Hahn, J. Gibbon, Y. Li, E.S. Parks, W.P. Parks and G.M. Shaw. 1988. Extensive variation of human immunodeficiency virus type-1 *in vivo*. Nature 334:440–444.

Saiki, R.K., S. Scharf, F. Faloona, K.B. Mullis, G.T. Horn, H.A. Erlich and N. Arnheim. 1985. Enzymatic amplification of β-globin genomic sequences and restriction site analysis for diagnosis of sickle cell anemia. Science 230:1350–1354.

Saitou, N. 1986. On the delta Q test of Templeton. Mol. Biol. Evol. 3:282–284.

Saitou, N and M. Nei. 1986. Polymorphism and evolution of influenza A virus genes. Mol. Biol. Evol. 3:57–74.

Saitou, N. and K. Omoto. 1987. Time and place of human origins from mtDNA data. Nature 327:288.

Sakoyama, Y., K.-J. Hong, S.M. Byun, H. Hisajima, S. Ueda, Y. Yaoita, H. Hayashida, T. Miyata and T. Honjo. 1987. Nucleotide sequences of immunoglobulin epsilon genes of chimpanzee and orangutan: DNA molecular clock and hominoid evolution. Proc. Natl. Acad. Sci. USA 84:1080–1084.

Salceda, V.M. 1977. Carga genetica en siete poblaciones naturales de *Drosophila melanogaster* (Meigen) de diferentes localidades de Mexico. Sobretiro de Agrociencia 28:47–52.

Sandler, L. and E. Novitski. 1957. Meiotic drive as an evolutionary force. Am. Nat. 91:105–110.

Sanger, F., S. Nicklen and A.R. Coulsen. 1977. DNA sequencing with chain–terminating inhibitors. Proc. Natl. Acad. Sci. USA 74:5463–5467.

Sarich, V.M. and A.C. Wilson. 1966. Quantitative immunochemistry and the evolution of primate albumins: Micro-complement fixation. Science 154:1563–1566.

Sarich, V.M. and A.C. Wilson. 1967. Immunological time scale for hominid evolution. Science 158:1200–1204.

Sasaki, T., T. Okazaki, M. Muramatsu and R. Kominami. 1987. Variation among mouse ribosomal RNA genes within and between chromosomes. Mol. Biol. Evol. 4:594–601.

Sawyer, S.A. D.E. Dykhuizen, R.F. DuBose, L. Green, T. Mutangadura-Mhlanga, D.F. Wolczyk and D.L. Hartl. 1987a. Distribution and abundance of insertion sequences among natural isolates of *Escherichia coli*. Genetics 115:51–63.

Sawyer, S.A., D.E. Dykhuizen and D.L Hartl. 1987b. Confidence interval for the number of selectively neutral amino acid polymorphisms. Proc. Natl. Acad. Sci. USA 84:6225–6228.

Sawyer, S.A. and D.L. Hartl. 1981. On the evolution of behavioral reproductive isolation: the Wallace effect. Theor. Pop. Biol. 19:261–273.

Sawyer, S.A. and D.L. Hartl. 1986. Distribution of transposable elements in prokaryotes. Theor. Pop. Biol. 30:1–17.

Sax, K. 1923. The association of size differences with seed coat pattern and pigmentation in *Phaseolus vulgarus*. Genetics 8:552–560.

Scavarda, N.J. and D.L. Hartl. 1984. Interspecific DNA transformation in *Drosophila*. Proc. Natl. Acad. Sci. USA 81:7515–7519.

Scavarda, N.J. and D.L. Hartl. 1987. Germ line abnormalities in *Drosophila simulans* transfected with the transposable *P* element. J. Genet. 66:1–15.

Schaeffer, S.W. and C.F. Aquadro. 1987. Nucleotide sequence of the *Adh* gene region of *Drosophila pseudoobscura*: Evolutionary change and evidence for an ancient gene duplication. Genetics 117:61–73.

Schaeffer, S.W., C.F. Aquadro and W.W. Anderson. 1987. Restriction-map variation in the *alcohol dehydrogenase* region of *Drosophila pseudoobscura*. Mol. Biol. Evol. 4:254–265.

Schaeffer, S.W., C.F. Aquadro and C.H. Langley. 1988. Restriction-map variation in the *Notch* region of *Drosophila melanogaster*. Mol. Biol. Evol. 5:30–40.

Scharf, S.J., G.T. Horn and H.A. Erlich. 1986. Direct cloning and sequence analysis of enzymatically amplified genomic sequences. Science 233:1076–1078.

Scharloo, W. 1971. Reproductive isolation by disruptive selection: Did it occur? Am. Nat. 105:83–86.

Scharloo, W. 1987. Constraints in selection response. pp. 125–150. *In* V. Loeschcke (ed.), *Genetic Constraints on Adaptive Evolution*. Springer-Verlag, Berlin.

Schenk, R.A. and R.C. Vrijenhoek. 1986. Spatial and temporal factors affecting coexistence among sexual and clonal forms of *Poeciliopsis*. Evolution 40:1060–1070.

Schlager, G. and M.M. Dickie. 1971. Natural mutations rates in the house mouse. Estimates for five specific loci and dominant mutations. Mut. Res. 11:89–96.

Schmalhausen, I.I. 1949. *Factors of Evolution: the Theory of Stabilizing Selection*. Blakiston, Philadelphia.

Schmid, C.W. and C.-K.J. Shen. 1985. The evolution of interspersed repetitive DNA sequences in mammals and other vertebrates. pp. 323–358. *In* R.J. MacIntyre (ed.), *Molecular Evolutionary Genetics*. Plenum, New York.

Schoen, D.J. and M.T. Clegg. 1986. Monte Carlo studies of plant mating system estimation models: The one-pollen parent and mixed mating models. Genetics 112:927–945.

Schull, W.J. and J.V. Neel. 1965. *The Effect of Inbreeding on Japanese Children*. Harper & Row, New York.

Schultz, R.J. 1969. Hybridization, unisexuality, and polyploidy in the teleost *Poeciliopsis* (Poeciliidae) and other vertebrates. Am. Nat. 103:605–619.

Seager, R.D. and F.J. Ayala. 1982. Chromosome interactions in *Drosophila melanogaster*. I. Viability studies. Genetics 102:467–483.

Seager, R.D., F.J. Ayala and R.W. Marks. 1982. Chromosome interactions in *Drosophila melanogaster*. II. Total fitness. Genetics 102:485–502.

Sederoff, R.R., C.S. Levings III, D.H. Timothy, W.W.L. Hu. 1981. Evolution of DNA sequence

organization in mitochondrial genomes of *Zea*. Proc. Natl. Acad. Sci. USA 78:5953–5957.

Selander, R.K. D.A. Caugant and T.S. Whittam. 1987. Genetic structure and variation in natural populations of *Escherichia coli*. pp. 1625–1648. *In* F.C. Neidhardt, J.L. Ingraham, L.K. Brooks, B. Magasanik, M. Schaechter and H.E. Umbarger (eds.), *Escherichia coli and Salmonella typhimurium: Cellular and Molecular Biology*, Vol. 2. American Society for Microbiology, Washington, D.C.

Selander, R.K., W.G. Hunt and W.G. Hunt. 1969. Protein polymorphism and genetic heterozygosity in two European subspecies of the house mouse. Evolution 23:379–390.

Selander, R.K. and B.R. Levin. 1980. Genetic diversity and structure in *Escherichia coli* populations. Science 210:545–547.

Selander, R.K., M.H. Smith, S.Y. Yang, W.E. Johnson and J.B. Gentry. 1971. Biochemical polymorphism and systematics in the genus *Peromyscus*. I. Variation in the old-field mouse (*Peromyscus polionotus*). Studies in Genetics VI. University of Texas Publ. 7103:49–90.

Service, P.M. and M.R. Rose. 1985. Genetic covariation among life history components: The effect of novel environments. Evolution 39:943–945.

Shaw, R.G. 1987. Maximum likelihood approaches applied to quantitative genetics of natural populations. Evolution 41:812–826.

Shen, S., J.L. Slightom and O. Smithies. 1981. A history of the fetal globin gene duplication. Cell 26:191–203.

Sheppard, P.M. 1959. The evolution of mimicry: A problem in ecology and genetics. Cold Spring Harbor Symp. Quant. Biol. 24:131–140.

Sherman, P.W. 1980. The limits of ground squirrel nepotism. pp. 505–544. In G.W. Barlow and J. Silverberg (eds.), *Sociobiology: Beyond Nature/Nurture?*. Westview Press, Boulder, CO.

Shields, J. 1962. *Monozygotic Twins Brought Up Apart and Brought Up Together*. Oxford University Press, London.

Shrimpton, A.E. and A. Robertson. 1988. The isolation of polygenic factors controlling bristle score in *Drosophila melanogaster*. II. Distribution of third chromosome bristle effects within chromosome sections. Genetics 118:445–459.

Sibley, C.G. and J.E. Ahlquist. 1981. The phylogeny and relationships of the ratite birds as indicated by DNA–DNA hybridization. pp. 301–335. *In* G.G.E. Scudder and J.L. Reveal (eds.), *Evolution Today*. (Proceedings of the Second International Congress on Systematics and Evolutionary Biology.) Carnegie Mellon, Pittsburgh.

Sibley, C.G. and J.E. Ahlquist. 1984. The phylogeny of the hominid primates as indicated by DNA–DNA hybridization. J. Mol. Evol. 20:2–15.

Sibley, C.G. and J.E. Ahlquist. 1986. Reconstructing bird phylogeny by comparing DNAs. Sci. Amer. 254:82–92.

Sibley, C.G. and J.E. Ahlquist. 1987. DNA hybridization evidence of hominoid phylogeny: Results from an expanded data set. J. Mol. Evol. 26:99–121.

Sick, K. 1965. Hemoglobin polymorphism of cod in the Baltic and the Danish belt sea. Hereditas 54:19–47.

Siegismund, H.R. and F.B. Christiansen. 1985. Selection components analysis of natural populations using population samples including mother–offspring combinations. III. Theor. Pop. Biol. 27:268–297.

Silver, L.M. 1985. Mouse *t* haplotypes. Annu. Rev. Genet. 19:179–208.

Silver, L.M. and D. Remis. 1987. Five of the nine genetically defined regions of mouse *t* haplotypes are involved in transmission ratio distortion. Genet. Res. 49:51–56.

Simmons, M.J. and J.F. Crow. 1977. Mutations affecting fitness in *Drosophila* populations. Annu. Rev. Genetics 11:49–78.

Simmons, M.J., C.R. Preston and W.R. Engels. 1980. Pleiotropic effects on fitness of mutations affecting viability in *Drosophila melanogaster*. Genetics 94:467–475.

Simmonds, N.W. 1977. Approximations for *i*, intensity of selection. Heredity 34:413–414.

Simon, M., J. Zieg, M. Silverman, G. Mandel and R. Doolittle. 1980. Phase variation: Evolution of a controlling element. Science 209:1370–1374.

Simpson, G.G. 1961. *Principles of Animal Taxonomy*. Columbia University Press, New York.

Sing, C.F. and J. Davignon. 1985. Role of apolipoprotein E polymorphism in determining normal plasma lipid and lipoprotein variation. Am. J. Hum. Genet. 37:268–285.

Singer, M.F. 1982. SINEs and LINEs: Highly repeated short and long interspersed sequences in mammalian genomes. Cell 28:433–434.

Singh, L., C. Phillips and K.W. Jones. 1984. The conserved nucleotide sequences of *Bkm*, which define *Sxr* in the mouse, are transcribed. Cell 36:111–120.

Singh, R.S., R.C. Lewontin and A.A. Felton. 1976. Genetic heterogeneity within electrophoretic "alleles" of *xanthine dehydrogenase* in *Drosophila pseudoobscura*. Genetics 84:609–629.

Singh, R.S. and L.R. Rhomberg. 1987a. A comprehensive study of genic variation in natural populations of *Drosophila melanogaster*. I. Estimates of gene flow from rare alleles. Genetics 115:313–322.

Singh, R.S. and L.R. Rhomberg. 1987b. A comprehensive study of genic variation in natural populations of *Drosophila melanogaster*. II. Estimates of heterozygosity and patterns of geographic differentiation. Genetics 117:255–271.

Skinner, S.W. 1982. Maternally inherited sex ratio in the parasitoid wasp *Nasonia vitripennis*. Science 215:1133–1134.

Skinner, S.W. 1985. *Son-killer*: A third extrachromosomal factor affecting the sex ratio in the parasitoid wasp *Nasonia* (= *Mormoniella*) *vitripennis*. Genetics 109:745–759.

Slagel, V., E. Flemington, V. Traina-Dorge, H. Bradshaw and P. Deininger. 1987. Clustering and subfamily relationships of the *Alu* family in the human genome. Mol. Biol. Evol. 4:19–29.

Slatkin, M. 1973. Gene flow and selection in a cline. Genetics 75:733–756.

Slatkin, M. 1977. Gene flow and genetic drift in a species subject to frequent local extinction. Theor. Pop. Biol. 12:253–262.

Slatkin, M. 1980. The distribution of mutant alleles in a subdivided population. Genetics 95:503–524.

Slatkin, M. 1981a. Estimating levels of gene flow in natural populations. Genetics 99:323–335.

Slatkin, M. 1981b. Populational heritability. Evolution 35:859–871.

Slatkin, M. 1985a. Rare alleles as indicators of gene flow. Evolution 39:53–65.

Slatkin, M. 1985b. Gene flow in natural populations. Annu. Rev. Ecol. Syst. 16:393–430.

Slatkin, M. 1986. Interchromosomal biased gene conversion, mutation and selection in a multigene family. Genetics 112:681–698.

Slatkin, M. 1987. The average number of sites separating DNA sequences drawn from a subdivided population. Theor. Pop. Biol. 32:42–49.

Slatkin, M. and T. Maruyama. 1975. Genetic drift in a cline. Genetics 81:209–222.

Slatkin, M. and N. Takahata. 1985. The average frequency of private alleles in a partially isolated population. Theor. Pop. Biol. 28:314–331.

Slightom, J.L., A.E. Blechl and O. Smithies. 1980. Human fetal $^{G}\gamma$ and $^{A}\gamma$ globin genes: Complete nucleotide sequences suggest that DNA can be exchanged between these duplicated genes. Cell 21:627–638.

Smith, C. 1975. Quantitative inheritance. pp. 382–441. *In* G. Fraser and O. Mayo (eds.), *Textbook of Human Genetics*. Blackwell Scientific, Oxford.

Smith, C.A.B. 1977. A note on genetic distance. Ann. Hum. Genet. 40:463–479.

Smith, R.L. 1974. *Ecology and Field Biology*, 2nd Ed. Harper & Row, New York.

Smouse, P.E. 1974. Likelihood analysis of recombinational disequilibrium in multiple-locus gametic frequencies. Genetics 76:557–565.

Smouse, P.E. 1986. The fitness consequences of multiple-locus heterozygosity under the multiplicative overdominance and inbreeding depression models. Evolution 40:946–957.

Smouse, P.E. and J.C. Long. 1988. A comparative *F*-statistic analysis of the genetic structure of human populations from lowland South America and highland New Guinea. pp. 32–47. *In* B.S. Weir, G. Eisen, M.M. Goodman and G. Namkoong (eds.), *Proceedings of the Second International Conference on Quantitative Genetics*. Sinauer Associates, Sunderland, MA.

Smouse, P.E. and J.V. Neel. 1977. Multivariate analysis of gametic disequilibrium in the Yanomama. Genetics 85:733–752.

Smouse, P.E., J.V. Neel and W. Liu. 1983. Multiple-locus departures from panmictic equilibrium within and between village gene pools of Amerindian tribes at different stages of agglomeration. Genetics 104:133–153.

Snyder, M., A.R. Fraser, J. LaRoche, K.E. Gartner-Kepkay and E. Zouros. 1987. Atypical mitochondrial DNA from a deep-sea scallop *Placopecten magellanicus*. Proc. Natl. Acad. Sci. USA 84:7595–7599.

Snyder, T.P. and F.J. Ayala. 1979. Frequency-dependent selection at the *PGM-1* locus of *Drosophila pseudoobscura*. Genetics 92:995–1003.

Sokal, R.R. and F.J. Rohlf. 1969. *Biometry*. W.H. Freeman, San Francisco.

Solignac, M. and M. Monnerot. 1986. Race formation, speciation and introgression within *Drosophila simulans*, *D. mauritiana*, and *D. sechellia* inferred from mitochondrial DNA analysis. Evolution 40:531–539.

Solignac, M., M. Monnerot and J.C. Mounolou. 1983. Mitochondrial DNA heteroplasmy in *Drosophila mauritiana*. Proc. Natl. Acad. Sci. USA 80:6942–6946.

Solignac, M., J. Genermont, M. Monnerot and J.C. Mounolou. 1984. Genetics of mitochondria in *Drosophila*: mtDNA inheritance in heteroplasmic strains of *D. mauritiana*. Mol. Gen. Genet. 197:183–188.

Solignac, M., J. Genermont, M. Monnerot and J.-C. Mounolou. 1987. *Drosophila* mitochondrial genetics: evolution of heteroplasmy through germ line cell divisions. Genetics 117:687–696.

Sourdis, J. and M. Nei. 1988. Relative efficiencies of the maximum parsimony and distance-matrix methods in obtaining the correct phylogenetic tree. Mol. Biol. Evol. 5:298–311.

Southern, E.M. 1975. Detection of specific sequences among DNA fragments separated by gel electrophoresis. J. Mol. Biol. 98:503–517.

Spencer, W.P. 1947. Mutations in wild populations of *Drosophila*. Advan. Genet. 1:359–402.

Spencer, H.G. and R.W. Marks. 1988. The maintenance of single-locus polymorphism. I. Numerical studies of a viability selection model. Genetics 120:605–613.

Spiess, E.B. 1968. Low frequency advantage in mating of *Drosophila pseudoobscura* karyotypes. Am. Nat. 102:363–379.

Spiess, E.B. 1970. Mating propensity and its genetic basis in *Drosophila*. Evol. Biol. 4 (Suppl.):315–380.

Spradling, A.C. and G.M. Rubin. 1981. *Drosophila* genome organization: Conserved and dynamics aspects. Annu. Rev. Genet. 15:219–264.

Sprague, G.F. 1978. Introductory remarks to the session on the history of hybrid corn. pp. 11–12. *In* D.B. Walden (ed.), *Maize Breeding and Genetics*. Wiley, New York.

Stacey, S.N., R.A. Lansman, H.W. Brock and T.A. Grigliatti. 1986. Distribution and conservation of mobile elements in the genus *Drosophila*. Mol. Biol. Evol. 3:522–534.

Stebbins, G.L. 1950. *Variation and Evolution in Plants*. Columbia University Press, New York.

Stebbins, G.L. 1976. Chromosome, DNA and plant evolution. Evol. Biol. 9:1–34.

Stebbins, G.L. and D.L. Hartl. 1988. Comparative evolution: Latent potentials for anagenic advance. Proc. Natl. Acad. Sci. USA 85:5141–5145.

Stephens, J.C. 1985. Statistical methods of DNA sequence analysis: Detection of intragenic recombination or gene conversion. Mol. Biol. Evol. 2:539–556.

Stephens, J.C. and M. Nei. 1985. Phylogenetic analysis of polymorphic DNA sequences at the *Adh* locus in *Drosophila melanogaster* and its sibling species. J. Mol. Evol. 22:289–300.

Stephenson, A.G. 1981. Flower and fruit abortion: Proximate causes and ultimate functions. Annu. Rev. Ecol. Syst. 12:253–279.

Stephenson, A.G. and J.A. Winsor. 1986. *Lotus corniculatus* regulates offspring quality through selective fruit abortion. Evolution 40:453–458.

Stewart, F.M. 1977. Appendix to P.A. Fuerst, R. Chakraborty and M. Nei, Statistical studies on protein polmorphism in natural populations. I. Distribution of single-locus heterozygosity. Genetics 86:455–483.

Stocker, B.A.D. 1949. Measurement of rate of mutation of flagellate antigenic phase in *Salmonella typhimurium*. J. Hygiene 47:398–412.

Strobeck, C. 1979a. Partial selfing and linkage: The effect of a heterotic locus on a neutral locus. Genetics 92:305–315.

Strobeck, C. 1979b. Haploid selection with n alleles in m niches. Am. Nat. 113:439–444.

Strobeck, C. and K. Morgan. 1978. The effect of intragenic recombination on the number of alleles in a finite population. Genetics 88:829–844.

Sudhof, T.C., J.L. Goldstein, M.S. Brown and D.W. Russell. 1985. The LDL receptor gene: A mosaic of exons shared with different proteins. Science 228:815–822.

Suzuki, H, N. Miyashita, K. Moriwaki, R. Kominami, M. Muramatsu, T. Kanehisa, F. Bonhomme, M.L. Petras, Z.-C. Yu and D.-Y. Lu. 1986. Evolutionary implication of heterogeneity of the nontranscribed spacer region of ribosomal DNA repeating units in various subspecies of *Mus musculus*. Mol. Biol. Evol. 3:126–137.

Sved, J.A. 1971. An estimate of heterosis in *Drosophila melanogaster*. Genet. Res. 18:97–105.

Sved, J.A. and F.J. Ayala. 1970. A population cage test for heterosis in *Drosophila pseudoobscura*. Genetics 66:97–113.

Sved, J.A. and B.H.D. Latter. 1977. Migration and mutation in stochastic models of gene frequency change. J. Math. Biol. 5:61–73.

Sved, J.A., T.E. Reed and W.F. Bodmer. 1967. The number of balanced polymorphisms that can be maintained in a natural population. Genetics 55:469–481.

Sylvester, J.E., D.A. Whiteman, R. Podolsky, J.M. Pozsgay, J. Respess and R.D. Schmickel. 1986. The human ribosomal RNA genes: Structure and organization of the complete repeating unit. Hum. Genet. 73:193–198.

Tachida, H. 1987. Differentiation of a multigene family between populations. Evolution 41:190–198.

Tajima, F. 1983. Evolutionary relationship of DNA sequences in finite populations. Genetics 105:437–460.

Takahata, N. 1981. Genetic variability and rate of gene substitution in a finite population under mutation and fluctuating selection. Genetics 98:427–440.

Takahata, N. 1983. Population genetics of extranuclear genomes under the neutral mutation hypothesis. Genet. Res. 42:235–256.

Takahata, N. 1984. A model of extranuclear genomes and the substitution rate under within-generation selection. Genet. Res. 44:109–116.

Takahata, N. 1987. On the overdispersed molecular clock. Genetics 116:169–179.

Takahata, N. 1988. More on the episodic clock. Genetics 118:387–388.

Takahata, N. and M. Kimura. 1979. Genetic variability maintained in a finite population under mutation and autocorrelated selection intensity. Proc. Natl. Acad. Sci. USA 76:5813–5817.

Takahata, N. and T. Maruyama. 1981. A mathematical model of extranuclear genes and the genetic variability maintained in a finite population. Genet. Res. 37:291–302.

Takahata, N. and M. Nei. 1984. F_{ST} and G_{ST} statistics in the finite island model. Genetics 107:501–504.

Takahata, N. and S.R. Palumbi. 1985. Extranuclear differentiation and gene flow in the finite island model. Genetics 109:441–457.

Tanksley, S.D., H. Medina-Filho and C.M. Rick. 1982. Use of naturally occurring enzyme variation to detect and map genes controlling quantitative traits in an interspecific backcross of tomato. Heredity 49:11–25.

Tautz, D., C. Tautz, D. Webb and G.A. Dover. 1987. Evolutionary divergence of promoters and spacers in the rDNA family of four *Drosophila* species: Implications for molecular coevolution in multigene families. J. Mol. Biol. 195:525–542.

Taylor, C.E. and V. Kekić. 1988. Sexual selection in a natural population of *Drosophila melanogaster*. Evolution 42:197–199.

Taylor, C.E. and J.R. Powell. 1977. Microgeographic differentiation of chromosomal and enzyme polymorphisms in *Drosophila persimilis*. Genetics 85:681–695.

Temin, R.G. and M. Marthas. 1984. Factors influencing the effect of segregation distortion in natural populations of *Drosophila melanogaster*. Genetics 107:375–393.

Templeton, A.R. 1979. Once again, why 300 species of Hawaiian *Drosophila*? Evolution 33:513–517.

Templeton, A. 1980. The theory of speciation via the founder principle. Genetics 94:1011–1038.

Templeton, A.R. 1983. Phylogenetic inference from restriction endonuclease cleavage site maps with particular reference to the evolution of humans and apes. Evolution 37:221–244.

Templeton, A.R. 1985. The phylogeny of the hominoid primates: A statistical analysis of the DNA–DNA hybridization data. Mol. Biol. Evol. 2:420–433.

Templeton, A.R. 1986. Further comments on the statistical analysis of DNA–DNA hybridization data. Mol. Biol. Evol. 3:290–295.

Templeton, A.R., E. Boerwinkle and C.F. Sing. 1987. A cladistic analysis of phenotypic associations with haplotypes inferred from restriction endonuclease mapping. I. Basic theory and an analysis of alcohol dehydrogenase activity in *Drosophila*. Genetics 117:343–351.

Templeton, A.R., T.J. Crease and F. Shah. 1985. The molecular through ecological genetics of abnormal abdomen in *Drosophila mercatorum*. I. Basic genetics. Genetics 111:805–818.

Thoday, J.M. 1961. Location of polygenes. Nature 191:368–370.

Thoday, J.M. 1979. Polygene mapping: Uses and limitations. pp. 219–234. *In* J.N. Thompson and J.M. Thoday (eds.), *Quantitative Genetic Variation*. Academic Press, New York.

Thoday, J.M. and J.B. Gibson. 1970. The probability of isolation by disruptive selection. Am. Nat. 104:219–230.

Thompson, E.A., S. Deeb, D. Walker, A.G. Motulsky. 1988. The detection of linkage disequilibrium between closely linked markers: RFLPs at the AI–CIII apolipoprotein genes. Am. J. Hum. Genetics 42:113–124.

Thompson, J.N. Jr. and J.M. Thoday. 1979. *Quantitative Genetic Variation*. Academic Press, London.

Thomson, G. 1977. The effect of a selected locus on linked neutral loci. Genetics 85:753–788.

Thomson, G. and W.F. Bodmer. 1977. The genetics of HLA and disease associations. pp. 545–564. *In* F.B. Christiansen and T.M. Fenchel (eds.), *Measuring Selection in Natural Populations*. Springer-Verlag, New York.

Thomson, G. and M.W. Feldman. 1974. Population genetics of modifiers of meiotic drive II. Linkage modification in the segregation distortion system. Theor. Pop. Biol. 5:155–162.

Thomson, G. and M.W. Feldman. 1975. Population genetics of modifiers of meiotic drive. IV. On the evolution of sex-ratio distortion. Theor. Pop. Biol. 8:202–211.

Thomson, G. and W. Klitz. 1987. Disequilibrium pattern analysis I. Theory. Genetics 116:623–632.

Tier, C. 1981. An analysis of neutral alleles and variable environment diffusion models. J. Math. Biol. 12:53–71.

Timofeeff-Ressovsky, H. 1935. Auslösnung von Vitalitatsmutationen durch Röntgenbestrahlung bei *Drosophila melanogaster*. Nachr. Ges. Wiss. Göttingen, Biol. N.F. 1:163–180.

Toong, Y.C., D.A. Schooley and F.C. Baker. 1988. Isolation of insect juvenile hormone III from a plant. Nature 333:170–171.

Tracey, M.L. and F.J. Ayala. 1974. Genetic load in natural populations: Is it compatible with the hypothesis that many polymorphisms are maintained by natural selection? Genetics 77:569–589.

Trippa, G., A. Loverre and A. Catamo. 1976. Thermostability studies for investigating non-electrophoretic polymorphic alleles in *Drosophila melanogaster*. Nature 260:42–44.

Trivers, R.L. and H. Hare. 1976. Haplodiploidy and the evolution of the social insects. Science 191:249–263.

Turelli, M. 1984. Heritable genetic variation via mutation-selection balance: Lerch's zeta meets the abdominal bristle. Theor. Pop. Biol. 25:138–193.

Turelli, M. 1985. Effects of pleiotropy on predictions concerning mutation-selection balance for polygenic traits. Genetics 111:165–195.

Turelli, M. 1986. Gaussian versus non-Gaussian models for polygenic mutation-selection balance. pp. 607–628. *In* S. Karlin and E. Nevo (eds.), *Evolutionary Processes and Theory*. Academic Press, New York.

Turelli, M. In press. Phenotypic evolution, constant covariances and the maintenance of additive variance. Evolution.

Turelli, M., J.H. Gillespie and R. Lande. 1988. Rate tests for selection on quantitative characters during macroevolution and microevolution. Evolution 42:1085–1089.

Turelli, M. and L.R. Ginzburg. 1983. Should individual fitness increase with heterozygosity? Genetics 104:191–209.

Turner, J.R.G. 1970. Changes in mean fitness under natural selection. in K. Kojima (ed.), *Mathematical Topics in Population Genetics*. Springer-Verlag, New York. pp. 32–78.

Turner, J.R.G. 1977. Butterfly mimicry: The genetical evolution of an adaptation. Evol. Biol. 10:163–206.

Turner, J.R.G. 1981. "Fundamental theorem" for two loci. Genetics 99:365–369.

Turner, J.R.G., M.S. Johnson and W.F. Eanes. 1979. Contrasted modes of evolution in the same genome: Allozymes and adaptive changes in *Heliconius*. Proc. Natl. Acad. Sci. USA 76:1924–1928.

Upholt, W.B. 1977. Estimation of DNA sequence divergence from comparison of restriction endonuclease digests. Nuc. Acids Res. 4:1257–1265.

Uyenoyama, M.K. 1984. Inbreeding and the evolution of altruism under kin selection: Effects on relatedness and group structure. Evolution 38:778–795.

Uyenoyama, M. 1985. Quantitative models of hybrid dysgenesis: Rapid evolution under transposition, extrachromosomal inheritance and fertility selection. Theor. Pop. Biol. 27:176–201.

Uyenoyama, M.K. and B.O. Bengtsson. 1979. Towards a genetic theory for the evolution of the sex ratio. Genetics 93:721–736.

Uyenoyama, M. and M.W. Feldman. 1978. The genetics of sex ratio distortion by cytoplasmic infection under maternal and contagious transmission: an epidemiological study. Theor. Pop. Biol. 14:471–497.

Uyenoyama, M.K. and M.W. Feldman. 1980. Theories of kin and group selection: A population genetics perspective. Theor. Pop. Biol. 17:380–414.

Uyenoyama, M.K. and M.W. Feldman. 1981. On relatedness and adaptive topography in kin selection. Theor. Pop. Biol. 19:87–123.

Uyenoyama, M.K., M.W. Feldman and L. Mueller. 1981. Population genetic theory of kin selection: Multiple alleles at one locus. Proc. Natl. Acad. Sci. USA 78:5036–5040.

Vadheim, C.M., J.I. Rotter, N.K. MacLaren, W.J. Riley and C.E. Anderson. 1986. Preferential transmission of diabetic alleles within the HLA gene complex. New Eng. J. Med. 315:1314–1318.

van Delden, W., A.C. Boerema and A. Kamping. 1978. The alcohol dehydrogenase polymorphism in populations of *Drosophila melanogaster*. I. Selection in different environments. Genetics 90:161–191.

van Noordwijk, A.J. and W. Scharloo. 1981. Inbreeding in an island population of the Great Tit. Evolution 35:674–688.

Van Valen, L. 1973. A new evolutionary law. Evolutionary Theory 1:1–30.

Via, S. and R. Lande. 1985. Genotype-environment interaction and the evolution of phenotypic plasticity. Evolution 39:505–522.

Voelker, R.A., C.C. Cockerham, F.M. Johnson, H.E. Schaffer, T. Mukai and L.E. Mettler. 1978. Inversions fail to account for allozyme clines. Genetics 88:515–527.

Voelker, R.A., H.E. Schaffer and T. Mukai. 1980. Spontaneous allozyme mutations in *Drosophila melanogaster*: rate of occurrence and nature of the mutants. Genetics 94:961–968.

Vrijenhoek, R.C. 1972. Genetic relationships of unisexual hybrid fishes to their progenitors using lactate dehydrogenase isozymes as gene markers (*Poeciliopsis*: Poeciliidae). Am. Nat. 106:754–766.

Vrijenhoek, R.C. 1978. Coexistence of clones in a heterogeneous environment. Science 199:549–552.

Wade, M.J. 1978. A critical review of models of group selection. Quart. Rev. Biol. 53:101–114.

Wade, M.J. 1982. The evolution of interference competition by individual, family, and group selection. Proc. Natl. Acad. Sci. USA 79:3575–3578.

Wade, M.J. 1985. Soft selection, hard selection, kin selection, and group selection. Am. Nat. 125:61–73.

Wade, M.J. and D.E. McCauley. 1980. Group selection: the phenotypic and genotypic differentiation of small populations. Evolution 34:799–812.

Wallace, B. 1958. The average effect of radiation-induced mutations on viability in *Drosophila melanogaster*. Evolution 12:532–552.

Wallace, B. 1968. *Topics in Population Genetics*. W.W. Norton, New York.

Walsh, J.B. 1985. How many processed pseudogenes are accumulated in a gene family? Genetics 110:345–364.

Walsh, J.B. 1986. Selection and biased gene conversion in a multigene family: consequences of interallelic bias and threshold selection. Genetics 112:699–716.

Walsh, J.B. 1987. Sequence-dependent gene conversion: can duplicated genes diverge fast enough to escape conversion? Genetics 117:543–557.

Waples, R.S. 1987. A multispecies approach to the analysis of gene flow in marine shore fishes. Evolution 41:385–400.

Watanabe, T. 1967. Infectious drug resistance. Sci. Am. 217:19–27.

Watanabe, T.K. 1979. A gene that rescues the lethal hybrids between *Drosophila melanogaster* and *D. simulans*. Japanese J. Genet. 54:325–331.

Watt, W.B. 1977. Adaptation at specific loci. I. Natural selection on phosphoglucose isomerase of *Colias* butterflies: Biochemical and population aspects. Genetics 87:177–194.

Watt, W.B. 1983. Adaptation at specific loci. II. Demographic and biochemical elements in the maintenance of the *Colias PGI* polymorphism. Genetics 103:691–724.

Watt, W.B., P.A. Carter and S.M. Blower. 1985. Adaptation at specific loci. IV. Differential mating success among glycolytic allozyme genotypes of *Colias* butterflies. Genetics 109:157–175.

Watt, W.B., R.C. Cassin and M.S. Swan. 1983. Adaptation at specific loci. III. Field behavior and survivorship differences among *Colias PGI* genotypes are predictable from in vitro biochemistry. Genetics 103:725–739.

Watterson, G.A. 1975. On the number of segregating sites in genetical models without recombination. Theor. Pop. Biol. 7:256–276.

Watterson, G.A. 1978. The homozygosity test of neutrality. Genetics 88:405–417.

Wayne, R.K., W.S. Modi and S.J. O'Brien. 1986. Morphological variability and asymmetry in the cheetah (*Acinonyx jubatus*), a genetically uniform species. Evolution 40:78–85.

Wehrhahn, C.F. 1975. The evolution of selectively similar electrophoretically detectable alleles in finite natural populations. Genetics 80:375–394.

Weinberg, W. 1908. On the demonstration of heredity in man. Translated by S.H. Boyer. 1963. *In Papers on Human Genetics*. Prentice-Hall, Englewood Cliffs, NJ.

Weir, B.S. 1970. Equilibrium under inbreeding and selection. Genetics 65:371–378.

Weir, B.S. 1979. Inferences about linkage disequilibrium. Biometrics 35:235–254.

Weir, B.S., R.W. Allard and A.L. Kahler. 1972. Analysis of complex allozyme polymorphisms in a barley population. Genetics 72:505–523.

Weir, B.S., A.H.D. Brown, D.R. Marshall. 1976. Testing for selective neutrality of electrophoretically detectable protein polymorphisms. Genetics 84:639–659.

Weir, B.S. and C.C. Cockerham. 1973. Mixed self and random mating at two loci. Genet. Res. 21:247–262.

Weir, B.S. and C.C. Cockerham. 1978. Testing hypotheses about linkage disequilibrium with multiple alleles. Genetics 88:633–642.

Weir, B.S. and C.C. Cockerham. 1979. Estimation of linkage disequilibrium in randomly mating populations. Heredity 42:105–111.

Weir, B.S. and C.C. Cockerham. 1984. Estimating *F* statistics for the analysis of population structure. Evolution 38:1358–1370.

Weir, B.S. and C.C. Cockerham. 1988. Analysis of linkage disequilibrium. *In* W.G. Hill and T.F.C. Mackay (eds.), *Festschrift for Alan Robertson*. Commonwealth Agricultural Bureau.

Wellauer, P and I.B. Dawid. 1979. Isolation and sequence organization of human ribosomal DNA. J. Mol. Biol. 128:289–303.

Weller, J.I. 1986. Maximum likelihood techniques for the mapping and analysis of quantitative trait loci with the aid of genetic markers. Biometrics 42:627–640.

Weller, J.I., M. Soller and T. Brody. 1988. Linkage analysis of quantitative traits in an interspecific cross of tomato (*Lycopersicon esculentum* × *Lycopersicon pimpinellifolium*) by means of genetic markers. Genetics 118:329–339.

Werren, J.H. 1983. Sex ratio evolution under local mate competition in a parasitic wasp. Evolution 37:116–124.

Werren, J.H., U. Nur and D. Eickbush. 1987. An extrachromosomal factor causing loss of paternal chromosomes. Nature 327:75–76.

Werren, J.H., S.W. Skinner and E.L. Charnov. 1981. Paternal inheritance of a *daughterless* sex ratio factor. Nature 293:467–468.

White, M.J.D. 1968. Models of speciation. Science 159:1065–1070.

White, M.J.D. 1969. Chromosomal rearrangements and speciation in insects. Annu. Rev. Genet. 3:75–98.

White, M.J.D. (ed.). 1974. *Genetic Mechanisms of Speciation in Insects*. Australia and New Zealand Book Co., Sydney.

White, M.J.D. 1978. *Modes of Speciation*. W.H. Freeman, San Francisco.

Whittam, T.S. 1981. Is a negative regression of Δp on p evidence of a stable polymorphism? Evolution 35:595–596.

Whittam, T.S., A.G. Clark, M. Stoneking, R.L. Cann and A.C. Wilson. 1986. Allelic variation in human mitochondrial genes based on patterns of restriction site polymorphisms. Proc. Natl. Acad. Sci. USA 83:9611–9615.

Whittam, T.S., H. Ochman and R.K. Selander. 1983. Multilocus genetic structure in natural populations of *Escherichia coli*. Proc. Natl. Acad. Sci. USA 80:1751–1755.

Widegren, B., U. Arnason and G. Akusjarvi. 1985. Characteristics of conserved 1,579-bp highly repetitive component in the killer whale, *Orcinus orca*. Mol. Biol. Evol. 2:411–419.

Wijsman, E.A. and L.L. Cavalli-Sforza. 1984. Migration and genetic population structure with special reference to humans. Annu. Rev. Ecol. Syst. 15:279–301.

Wilkinson, G.S. 1987. Equilibrium analysis of sexual selection in *Drosophila melanogaster*. Evolution 41:11–21.

Williams, S.M., R. DeSalle and C. Strobeck. 1985. Homogenization of geographical variants at the nontranscribed spacer of rDNA in *Drosophila mercatorum*. Mol. Biol. Evol. 2:338–346.

Williams, S.M., G.R. Furnier, E. Fuog and C. Strobeck. 1987. Evolution of the ribosomal DNA spacers of *Drosophila melanogaster*: Different patterns of variation on X and Y chromosomes. Genetics 116:225–232.

Wilson, A.C., G.L. Bush, S.M. Case and M.-C. King. 1975. Social structuring of mammalian populations and rate of chromosomal evolution. Proc. Natl. Acad. Sci. USA 72:5061–5065.

Wilson, A.C., H. Ochman and E.M. Prager. 1987. Molecular time scale for evolution. Trends in Genet. 3:241–247.

Wilson, D.S. 1983. The group selection controversy: history and current status. Annu. Rev. Ecol. Syst. 14:159–188.

Wilson, D.S. and M. Turelli. 1986. Stable underdominance and the evolutionary invasion of empty niches. Am. Nat. 127:835–850.

Wilson, E.O. 1971. *The Insect Societies*. Harvard University Press, Cambridge, MA.

Wilson, E.O. 1975. *Sociobiology: The New Synthesis*. Harvard University Press, Cambridge, MA.

Wilson, S.R., J.G. Oakeshott, J.B. Gibson and P.R. Anderson. 1982. Measuring selection coefficients affecting the *alcohol dehydrogenase* polymorphism in *Drosophila melanogaster*. Genetics 100:113–126.

Wilton, A.N., C.C. Laurie-Ahlberg, T.H. Emigh and J.W. Curtsinger. 1982. Naturally occurring enzyme activity variation in *Drosophila melanogaster*. II. Relationship among enzymes. Genetics 102: 207–221.

Woese, C.R. 1981. Archaebacteria. Sci. Amer. 244(6):98–122.

Wolfe, K.H., W.-H. Li and P.M. Sharp. 1987. Rates of nucleotide substitution vary greatly among plant mitochondrial, chloroplast and nuclear DNAs. Proc. Natl. Acad. Sci. USA 84:9054–9058.

Wolstenholme, G.E.W. and M. O'Connor (eds.). 1969. *Bacterial Episomes and Plasmids*. Little, Brown and Co., Boston.

Woo, S.L.C., A.S. Lidsky, F. Guttler, T. Chandra and K.J.H. Robson. 1983. Cloned human phenylalanine hydroxylase gene allows prenatal diagnosis and carrier detection of classical phenylketonuria. Nature 306:151–155.

Workman, P.L. 1964. The maintenance of heterozygosity by partial negative assortative mating. Genetics 50:1369–1382.

Workman, P.L and S.K. Jain. 1966. Zygotic selection under mixed random mating and self-fertilization: Theory and problems of estimation. Genetics 54:159–171.

Worton, R.G., J. Sutherland, J.E. Sylvester, H.F. Willard, S. Bodrug, I. Dube, C. Duff, V. Kean, P.N. Ray and R.D. Schmickel. 1988. Human ribosomal RNA genes: Orientation of the tandem array and conservation of the 5′ end. Science 239:64–68.

Wright, J.E., K. Johnson, A. Hollister and B. May. 1983. Meiotic models to explain classical linkage, pseudolinkage, and chromosome pairing in tetraploid derivative salmonid genomes. Isozymes: Curr. Top. Biol. Med. Res. 10:239–260.

Wright, S. 1921. Systems of mating, I-V. Genetics 6:111–178.

Wright, S. 1931. Evolution in Mendelian populations. Genetics 16:97–159.

Wright, S. 1932. The roles of mutation, inbreeding, crossbreeding and selection in evolution. Proc. 6th Int. Cong. Genet. 1:356–366.

Wright, S. 1938. Size of population and breeding structure in relation to evolution. Science 87:430–431.

Wright, S. 1967. "Surfaces" of selective value. Proc. Natl. Acad. Sci. USA 58:165–172.

Wright, S. 1968. *Evolution and the Genetics of Populations. Vol. 1: Genetic and Biometric Foundations*. University of Chicago Press, Chicago.

Wright, S. 1969. *Evolution and the Genetics of Populations. Vol. 2: The Theory of Gene Frequencies*. University of Chicago Press, Chicago.

Wright, S. 1977. *Evolution and the Genetics of Populations. Vol. 3: Experimental Results and Evolutionary Deductions*. University of Chicago Press, Chicago.

Wright, S. 1978. *Evolution and the Genetics of Populations. Vol. 4: Variability within and among Natural Populations*. University of Chicago Press, Chicago.

Wright, S. 1988. Surfaces of selective value revisited. Am. Nat. 131:115–123.

Wu, C.-I. 1983a. The fate of autosomal modifiers of the sex-ratio trait in *Drosophila* and other sex-linked meiotic drive systems. Theor. Pop. Biol. 24:107–120.

Wu, C.-I. 1983b. Virility deficiency and the sex-ratio trait in *Drosophila pseudoobscura* I. Sperm displacement and sexual selection. Genetics 105:651–662.

Wu, C.-I. 1983c. Virility deficiency and the sex–ratio trait in *Drosophila pseudoobscura* II. Multiple mating and overall virility selection. Genetics 105:663–679.

Wu, C.-I. and W.-H. Li. 1985. Evidence for higher rates of nucleotide substitution in rodents than in man. Proc. Natl. Acad. Sci USA 82:1741–1745.

Wu, C.-I. and N. Maeda. 1987. Inequality in mutation rates of the two strands of DNA. Nature 327:169–170.

Wynne-Edwards, V.C. 1962. *Animal Dispersion in Relation to Social Behaviour*. Oliver and Boyd, Edinburgh.

Yamazaki, T. 1971. Measurement of fitness at the *esterase-5* locus in *Drosophila pseudoobscura*. Genetics 67:579–603.

Yokoyama, S., L. Chung and T. Gojobori. 1988. Molecular evolution of the human immunodeficiency and related viruses. Mol. Biol. Evol. 5:237–251.

Young, S.S.Y. 1967. A proposition on the population dynamics of the sterile *t* alleles in the house mouse. Evolution 21:190–198.

Young, S.S.Y. and H. Weiler. 1960. Selection for two correlated traits by independent culling levels. J. Genet. 57:329–338.

Youssoufian, H., S.E. Antonarakis, W. Bell, A.M. Griffin and H.H. Kazazian. 1988. Nonsense and missense mutations in hemophilia A: Estimate of the relative mutation rate at CG dinucleotides. Am. J. Hum. Genet. 42:718–725.

Yunis, J.J. and O. Prakash. 1982. The origin of man: A chromosomal pictorial legacy. Science 215:1525–1530.

Zlotogora, J., M. Zeigler and G. Bach. 1988. Selection in favor of lysosomal storage disorders? Am. J. Hum. Genet. 42:271–273.

Zouros, E. 1976. The distribution of enzyme and inversion polymorphism over the genome of *Drosophila*: Evidence against balancing selection. Genetics 83:169–179.

Zouros, E., S.M. Singh and H.E. Miles. 1980. Growth rate in oysters: An overdominant phenotype and its possible explanations. Evolution 34:856–867.

Zuckerkandl, E. and L. Pauling. 1965. Evolutionary divergence and convergence in proteins. pp. 97–166. *In* V. Bryson and H.J. Vogel (eds.), *Evolving Genes and Proteins*. Academic Press, New York.

Author Index

Abramowitz, M., 128
Aguadé, M., 336, 337
Ahlquist, J.E., 330, 345, 366
Allard, R.W., 19, 51
Allison, A.C., 169
Altenberg, L., 217
Amato, G.D., 223
Anderson, E.S., 583
Anderson, W.W., 175, 184, 185, 221, 517, 533
Andjelkovic, M., 223
Antonovics, J., 548, 549, 550, 552, 553
Anxolabéhère, D., 221, 416
Aoki, K.-I., 314, 315, 326
Aquadro, C.F., 27, 335, 337, 338, 342, 343
Arnheim, N.M., 395
Arnold, J., 496, 497
Arnold, S.J., 195
Ashburner, M., 343
Asmussen, M.A., 425
Atchley, W.R., 475, 492, 496
Avise, J.C., 90, 422, 423, 597
Avni, H., 219
Axelrod, R., 568
Ayala, F.J., 122, 126, 185, 221, 252, 522, 523

Baird, M., 140
Baker, W.K., 51
Baltimore, D., 102
Barker, J.S.F., 216, 219, 530
Barnes, P.T., 542
Barton, N.H., 322, 494
Bates, H.W., 570
Beckmann, J.S., 507
Beecham, E., 496
Bell, G., 217
Bengtsson, B.O., 556
Benjamin, D.C., 347
Bennett, J.H., 57, 182, 267
Benveniste, R.E., 362, 416–420, 584
Berg, D.E., 584
Bermingham, E., 426
Bernstein, S.C., 23
Betts, P.W., 395
Beverly, S.M., 516, 593
Biémont, C., 415
Bijlsma, R., 224, 507
Binet, F.E., 267
Bingham, P.M., 415
Birky, C.W., 351, 425
Birley, A.J., 25
Bishop, J.M., 420, 579, 581
Bodmer, W.F., 54, 169, 171, 188, 309
Bodmer, M., 343
Boggs, C.L., 195
Boller, E.F., 596
Bonhomme, F., 24
Bonnell, M.L., 84
Boorman, S.A., 561
Boursot, P., 426
Boussy, I.A., 416
Boyd, R., 569
Bradshaw, A.D., 548, 550–552
Bregliano, J.-C., 416
Briggs, G.E., 539, 541
Britten, R.J., 362
Brittnacher, J.G., 191
Brookfield, J.F.V., 408
Brower, J.V.Z., 572
Brower, L.P., 573, 575
Brown, A.H.D., 114–117, 220
Brown, C.J., 184
Brown, S.W., 566
Brown, W.M., 421–423
Brownell, E.M., 388
Bryan, G.J., 112
Bryant, E.H., 186, 595
Bryga, H., 537
Bucheton, A.R., 416
Bull, J.J., 558, 559, 560
Bulmer, M.G., 480, 494
Bumpus, H.C., 497
Bundgaard, J., 193, 222, 223
Buonagurio, D.A., 365
Burgman, M.A., 379
Buri, P., 66, 69, 78, 79
Burns, J.A., 540, 541, 542
Burr, B., 507
Burton, R.S., 542, 544, 547, 548
Bush, G.L., 591, 592, 595–597, 602
Bush, R.M., 555

Caisse, M., 553
Campbell, H.M., 185, 580, 582, 583
Cann, R.L., 92, 424
Cannings, C., 182
Cantor, C.R., 357
Carmelli, D., 215
Carson, H.L., 516, 591–593, 595
Carter, L.O., 483
Castle, W.E., 32, 485
Cavalli-Sforza, L.L., 169–171, 306, 309, 565
Cavender, D.E., 228
Cavener, D.R., 509, 511
Chakraborty, R. 84, 137, 138, 139
Chammah, A.M., 568
Champion, A.B., 347
Chao, L.C., 411, 412
Chapman, R.W., 425
Chappell, M.A., 546
Charlesworth, B., 386, 399, 408, 504, 517, 533, 534, 537, 565, 575, 576
Charlesworth, D., 408, 575, 576
Charnov, E.L., 560
Chesser, R.K., 297, 322
Christiansen, F.B., 221–227, 292, 529
Clare, M.J., 537
Clark, A.G., 140, 157, 190, 218, 221–223, 376, 425, 426, 507, 524, 538, 542
Clarke, B., 474, 595
Clayton, G.A., 446, 499
Clegg, M.T., 19, 51, 114, 115, 116, 117, 218, 260, 424, 509, 511
Cleve, H., 595
Clowes, R.C., 580
Cluster, P.D., 395

Cockerham, C.C., 55, 56, 57, 219, 227, 264, 266, 271, 293, 297–300, 322, 487, 499
Coen, E.S., 393
Cohan, F.M., 509, 510
Cohn, V.H., 343
Cook, L.M., 579, 581
Cook, R.D., 529
Cook, S.C.A., 550
Cotterman, C.W., 469
Coyne, J.A., 24, 496, 600, 601
Creed, E.R., 579
Crick, F., 400, 426
Cross, S.R.H., 25
Crow, J.F., 71, 80, 82, 88, 121, 199, 216, 298, 303, 313, 314, 315, 326, 363, 469, 502, 555
Crozier, R.H., 566
Crumpacker, D.W., 253
Curie-Cohen, M., 297
Curtis, S.E., 424
Curtsinger, J.H., 181, 185, 227, 558

Daniels, G.R., 403
Darnell, J., 102
Darwin. C., 147–149, 194, 195, 197, 198, 349, 430, 497
Davies, J., 582, 584
Davignon, J., 507, 508
Davis, P.S., 409
Dawid, I.B., 395
Dawood, M.M., 185
Dean, A.M., 541
DeJong, G., 223, 507
Deininger, P.L., 403
Delarbre, C.Y., 193
Demers, G.W., 403
Dempster, E.R., 525
Dempster, A.P., 42
Denniston, C., 303
Densmore, L.D., 426
DeSalle, R.J., 366, 547
Dice, L.R., 90
Dickerson, R.E., 169
Dickie, M.M., 107
DiLella, A.G., 543
DiMichele, L., 546
Doane, W.W., 223, 507
Dobzhansky, Th., 99, 100, 162, 175, 220, 253, 473, 517, 600
Donis-Keller, H., 14, 42
Doolittle, F.W., 397, 400
Dover, G., 400
Dozy, A.M., 339
DuBose, R.F., 341
DuMouchel, W.H., 221
Dudley, J.W., 442
Dunn, L.C., 191, 193
Dutton, F.L., 394
Dykhuizen, D.E., 150, 183, 224, 541

Eanes, W.F., 182, 224
East, E.M., 274
Easteal, S., 363
Easterby, J., 542
Edwards, A.W.F., 303, 304
Edwards, M.D., 506
Eggleston, W.B., 416
Ehrlich, P.R., 569
Ehrman, L., 184
Elton, C., 85
Elston, R., 507
Endow, S.A., 394
Enfield, F.D., 436, 446
Engelke, D.R., 343
Engels, W.R., 25, 332, 333, 415, 416
Eshel, I., 216, 557, 566
Ewens, W.J., 69, 82, 126–128, 130, 139–143, 150, 179, 213, 228, 333

Falconer, D.S., 446, 448, 456, 476, 479, 490
Falk, C.T., 276
Fedoroff, N.V., 407
Feldman, M.W., 110, 174, 186, 190, 215–217, 222, 227, 277, 292, 471, 524, 544, 547, 557, 558, 560, 563, 565, 569
Felsenstein, J., 82, 88, 216, 379
Ferrell, R.E., 136, 139, 142
Ferris, S.D., 29, 597
Fisher, R.A., 5, 61, 66, 69–71, 80, 164, 194, 207, 435, 504, 533, 555
Fitch, W.M., 364, 378, 379
Fleming, W.H., 502, 503
Flint, J., 391, 392
Foley, P., 504
Ford, H.A., 571
Foster, G.G., 163, 165
Frankham, R., 393
Franklin, I., 218
Frischauf, A.-M., 193
Frydenberg, O., 221, 224–227
Fryer, G., 595
Fuerst, P.A., 136
Futuyma, D.J., 570

Gaffney, B., 553
Galton, F., 435
Ganetzky, B., 191
Gardner, P., 584
Gause, G.F., 589
Geer, B.W., 542
Geis, I., 169
Geldermann, H., 507
George, J.T., 507
Gibbs, K.L., 497
Gibson, J.B., 590
Giddings, L.V., 366
Giesbrecht, F.G., 489
Giesel, J.D., 537
Gilbert, W., 340, 341, 398
Gillespie, J.H., 24, 110, 143, 363, 364, 529
Gillings, M.R., 394
Gingerich, P.D., 364
Ginzburg, L.R., 165, 218, 554
Gliddon, C., 529
Glover, D.M., 393
Goeddel, D.V., 354
Gojobori, T., 395
Golding, G.B., 117, 414
Golenberg, E.M., 165
Goodman, M., 389
Gotoh, O., 335
Gould, S.J., 517, 539
Graf, J.D., 509, 510
Grant, P.R., 497, 498, 553
Gray, P.W., 354
Greaves, J.H., 199
Green, L., 42
Griffiths, D.J., 553
Grun, P., 426
Gupta, A.P., 472, 473
Gusella, J.F., 28
Gwynne, D.T., 195

Hadeler, K.P., 189
Hairston, N.G., 535
Haldane, J.B.S., 118, 181, 529, 533, 539, 541, 563, 578
Hale, L.R., 426
Hall, B.G., 395
Hamilton, W.D., 535, 556, 558, 565, 567, 568
Hamrick, J.L., 19
Hardies, S.C., 389, 403–406
Hardin, G., 589
Hardison, R.C., 389
Hardy, G.H., 32
Hare, H., 567
Harris, H., 9, 19, 21
Harris, S., 363
Harris, T., 497
Harrison, R.G., 597
Hartl, D.L., 92, 150, 164, 182, 183, 191, 222, 408, 412–414, 416, 515, 516, 529, 542, 566, 591
Hastings, A., 215, 216, 218, 219, 504

Hauswirth, W.W., 426
Hayashida, H., 365
Haymer, D.S., 164, 222
Head, G., 559
Hedrick, P.W., 54, 57, 182, 186, 216, 219, 222, 228, 303
Heinrich, R., 542
Herre, E.A., 558
Hickey, D.A., 185, 223, 507
Hilbish, T.J., 544, 545
Hill, W.G., 55, 56, 82, 88, 219, 499, 500
Hinds, K.R., 398
Hiraizumi, Y., 191, 199, 538
Hoekstra, R.F., 528
Hoffman, A.A., 595
Hogg, J.T., 194
Hogness, D.S., 393
Holgate, P., 84
Holland, J.K., 365
Hollingsworth, T.H., 85
Holmquist, R., 366
Holsinger, K.E., 110, 216
Hood, L., 397
Hopkinson, D.A., 9, 19
Howard, W.B., 90
Hubby, J.L., 9, 19
Hudson, R.R., 130, 131, 230, 377, 383
Huettel, M.D., 596
Hughes, A.L., 377
Hughes, D.M., 538
Hunkapiller, T., 397

Ikemura, T., 372
Iles, T.D., 595
Inouye, S. 23
Ives, P.T., 200

Jacquard, A., 479
Jaenike, J., 549, 550
Jain, S.K., 269, 548, 550, 551, 552
Jayakar, S.D., 181, 507, 529, 566
Johannsen, W., 441, 444
Johnson, F.M., 10
Johnson, K.R., 603
Jones, J.S., 474
Jukes, T.H., 136, 357, 362
Jungen, H., 164, 222

Kacser, H., 540, 541, 542
Kallman, K.D., 558
Kan, Y.W., 339
Kaneshiro, K.Y., 516, 595
Kaplan, N.L., 334, 335, 383, 408
Karlin, S., 127, 128, 174, 197, 215, 219, 275, 277, 294, 347, 528, 557, 565
Karn, M.N., 497
Kazazian, H.H., 339, 403
Keightley, P.D., 542
Keith, L.E., 542
Keith, T.P., 24, 127, 141, 228, 376
Keller, E.F., 407
Kekić, V., 198, 498
Kempthorne, O., 479, 533
Kerr, W.E., 79, 566, 567
Kerster, H.W., 551
Kessler, S., 589
Kettlewell, H.B.D., 577, 578
Keyfitz, N., 531
Kidd, K.K., 543
Kidwell, J.F., 180
Kidwell, M.G., 415, 416
Kilmer, W.L., 561
Kim, J., 379
Kimura, M., 71–73, 80, 82, 88, 124, 133, 135, 136, 143, 269, 313, 316–320, 349, 350, 359–361, 364, 371, 469, 502, 529, 555
King, C.E., 533
King, J.L., 136, 362
Kingman, J.F.C., 90, 92, 135, 172, 173, 175, 176, 503
Kirkpatrick, M., 195
Kiyasu, P.K., 416
Klarenberg, A.J., 51
Kleckner, N., 584, 585
Kluge, R., 507
Knight, G.R., 590
Koch, R.E., 117
Koehn, R.K., 544, 545, 553, 554
Kohn, L.A., 496
Kojima, K., 24, 215
Komma, D.J., 394
Kreitman, M., 30, 229, 336, 337, 339, 340, 342, 372, 377, 382, 600, 601
Kricker, M. 395
Krider, H.M., 394
Kubitschek, H.E., 183

Laipis, P.J., 426
Lake, J.A., 366
Lande, R., 195, 476, 487, 488, 493, 496, 497, 499–503
Lander, E.S., 42
Landsmann, J., 360
Langley, C.H., 24, 51, 335, 337, 338, 364, 408
Lansman, R.A., 414
LaPorte, D.C., 542
Laski, F.A., 414
Latter, B.D.H., 125, 314
Laurie-Ahlberg, C.C., 7, 491, 542
Ledig, F.T., 553, 555
Lee, G.M., 603
Lees, D.R., 579
Leigh, E.G., 561
Leigh Brown, A.J., 24
Lenski, R.E., 586, 587
Lerner, I.M., 489
Lessard, S., 557
Levene, H., 524, 527, 528
Levin, B.R., 21, 217, 561, 586, 587
Levin, D.A., 19, 263, 296, 297, 551, 597
Levings, C.S., 426
Levitt, P.R., 561, 566
Levy, M., 597
Lewis, S.E., 10
Lewontin, R.C., 8, 9, 19, 51, 52, 174, 184, 185, 191, 193, 215, 218, 227, 306, 471, 472, 473, 517, 539, 562
Li, C.C., 32, 276
Li, W.-H., 114, 116, 117, 289, 291, 332, 334, 357, 363, 368, 369, 373–375, 395
Liberman, U., 110, 189, 216, 217, 219
Littman, G.W., 398
Lodish, H., 102
Lofsvold, D., 496
Long, J.C., 299–301
Lorberbaum, J.P., 569
Luckinbill, L.S., 537
Lyckegaard, E.M.S., 426
Lynch, M., 499, 500, 530

MacCluer, J.W., 242
Mackay, T.F.C., 409, 433, 445, 446
MacNair, M., 550
Maeda, N., 117, 384
Majors, J., 16
Mandel, S.P.H., 174, 181
Mange, A.P., 298
Mangelsdorf, A.J., 274
Manly, B.F.J., 142, 149
Margoliash, E., 378
Margot, J.B., 346, 389
Marinkovic, D., 19
Markow, T.A., 186
Marks, J., 389
Marks, R.W., 175, 524
Marthas, M., 191
Maruyama, T., 316–318
Matessi, C., 565, 566
Maxam, A.M., 340, 341

Maxson, L.R., 347–349
Maxson, R.D., 348, 349
Maynard Smith, J., 448, 528, 569, 591
Mayr, E., 589, 591
Mazer, S.J., 477
McAllister, M.K., 563
McCarrey, J.R., 404
McCauley, D.E., 562
McChesney, F., 524
McClintock, B., 407
McConnel, T.J., 367
McGregor, J., 127, 128
McKenzie, J.A., 474
McLellan, T., 23
Medawar, P.B., 535
Menten, M.L., 539
Meynell, G.G., 580
Michaelis, L., 539, 540
Michod, R.E., 217, 566
Miklos, G.L.G., 399, 400, 401
Miller, R.H., 372
Mitchell-Olds, T., 471, 476, 497
Mitton, J.B., 553
Möller, A.P., 198
Monnerot, M., 426
Montgomery, E.A., 408
Moran, P.A.P., 87, 88, 215
Morgan, K., 134
Morse, B., 403
Morton, N.E., 254, 256
Mourant, A.E., 37, 38, 40, 41
Mueller, L.D., 89, 522, 523, 530, 537
Mukai, T., 18, 51, 52, 55, 118–120, 223, 250, 252, 253
Muller, H.J., 104
Müller, F., 570
Murray, J., 222, 595

Nagylaki, T., 206, 386
Nakajima, K.U., 365
Nakamura, Y., 28
Nasrallah, J.B., 275
Nathans, J., 392
Neal, N.P., 253
Needleman, S.B., 345
Neel, J.V., 84, 108, 109, 199, 254
Nei, M., 84, 137, 216, 289, 291, 293, 297, 298, 302, 303, 305, 306, 314, 322, 324, 332, 334, 335, 365, 366, 377, 379, 380, 382
Nevo, E., 19, 20, 595
Newcombe, H.B., 39
Nichols, B.P., 358
Nicholson, M., 85
Nilsson-Ehle, H., 450
Nordskog, A.W., 489
Norton, H.T.J., 533, 534
Nöthel, H., 110
Novick, A., 112
Nunney, L., 185

O'Brien, S.J., 23, 84, 462
O'Connor, M., 580
O'Donald. P., 194, 197, 198
Oertel, C.R., 57
Ohnishi, O., 121
Ohta, T., 71, 135, 136, 269, 364, 382, 386–388
Omoto, K. 424
Orgel, L.E., 400
Orkin, S.H., 339
Oster, G., 567
Østergaard, H., 227

Palmer, J.D., 424
Palumbi, S.R., 316
Pantazidis, A.C., 604
Partridge, L., 498
Pauling, L., 360
Peck, J.R., 569
Peetz, E.W., 374
Penrose, L.S., 187, 497
Perutz, M.F., 5
Petit, C., 184
Pichersky, E., 392
Pirchner, F., 470
Place, A.R., 542
Platt, A.P., 573
Pollak, E., 190, 533
Porteous, J.W., 541
Powell, J.R., 42, 223, 422, 590, 595
Powers, D.A., 544, 546
Prakash, S., 19, 52, 598, 599
Press, W.H., 38
Price, T.D., 498
Prout, T., 162, 175, 191, 193, 221, 524, 529

Race, R.R., 54
Ramshaw, J.A.M., 136
Rand, D.M., 426
Raper, J., 197
Rapoport, A., 542, 568
Raven, P.H., 569
Rawlings, J.O., 271
Reed, T.E., 199
Reeder, R.H., 395
Remis, D., 192
Reznick, D.N., 537
Rhomberg, L.R., 305, 320, 321, 322
Rice, W.R., 559
Richardson, R.H., 516
Richmond, R.C., 19, 548
Riska, B., 474
Risko, K., 335
Robertson, A., 162, 463, 488, 499, 524
Robinson, H.F., 470
Rohlf, F.J., 379, 438
Roitberg, B.D., 563
Rose, M.R., 504, 537, 538
Roughgarden, J., 517, 520, 521
Royden, C.S., 339
Rubin, G.M., 414
Ruddle, F.H., 19
Russell, W.L., 107
Rutledge, J.J., 476
Rutowski, R.L., 195
Ruvolo, M., 366

Saag, M.S., 366
Saiki, R.K., 343
Saitou, N., 365, 366, 424
Sakoyama, Y., 363
Salceda, V.M., 200
Sanger, F., 54, 341
Sapienza, C., 400
Sarich, V.M., 347, 348
Sawyer, S.A., 229, 376, 408, 410, 411, 591
Sax, K., 505
Scavarda, N.J., 416
Schaeffer, S.W., 336, 338, 343
Scharf, S.J., 343
Scharloo, W., 51, 223, 241, 242, 448, 450, 507, 590
Schenk, R.A., 603
Schlager, G., 107
Schmalhausen, I.I., 472
Schmid, C.W., 402, 404
Schull, W.J., 254
Schultz, R.J., 603
Seager, R.D., 221
Selander, R.K., 19, 21, 24, 84
Service, P.M., 537
Seton, E.T., 85
Shaw, R.G., 478, 497
Shen, M.W., 389, 390, 402, 404
Sheppard, P.M., 570
Sherman, P.W., 566
Shields, J., 481
Shrimpton, A.E., 488
Sibley, C.G., 330, 345, 366
Sick, K., 285, 286, 287
Siegismund, H.R., 227
Silver, L.M., 191, 192
Simmonds, N.W., 483
Simmons, M.J., 121, 199, 216, 538
Simon, M., 114

Simpson, G.G., 588
Sing, C.F., 507, 508
Singh, L., 401
Singh, R.S., 24, 305, 320–322, 426
Skinner, S.W., 560
Slagel, V., 403
Slatkin, M., 316, 317, 319, 320, 322, 388, 562, 570
Slightom, J.L., 389
Smith, C., 471, 481, 484
Smith, C.A.B., 303
Smith, D.I., 582
Smith, R.L., 561
Smith, T.F., 366
Smithies, O., 384
Smouse, P.E., 57, 220, 300, 555
Snyder, M., 421
Snyder, T.P., 185
Solignac, M., 426
Soller, M., 507
Sondhi, K.C., 448
Sourdis, J., 379
Southern, E.M., 12, 13, 331
Spassky, B., 473
Spencer, H.G., 175
Spencer, W.P., 8
Spiess, E.B., 184
Spradling, A.C., 414
Sprague, G.F., 261
Stacey, S.N., 415
Stebbins, G.L., 96, 515, 516, 596
Stegun, I.A., 128
Stephens, J.C., 380, 382
Stephenson, A.G., 560
Stewart, F.M., 140, 141
Stocker, B.A.D., 114
Strickberger, M.W., 185
Strobeck, C., 134, 271, 529
Südhof, T.C., 397
Sved, J.A., 122, 125, 221, 252

Tachida, H., 388, 499
Tajima, F., 132, 135, 335, 366
Takahata, N., 314, 316, 322, 364, 425, 529
Tanksley, S.D., 507
Tautz, D., 394
Taylor, C.E., 42, 198, 498
Taylor, H.M., 294
Temin, R.G., 191, 199
Templeton, A.R., 366, 516, 547, 591, 592
Thoday, J.M., 505, 590
Thomas, E.A.C., 191, 404, 569
Thompson, E.A., 55, 84
Thomson, G., 54, 186, 216, 228, 557, 558
Tier, C., 529
Timofeeff-Ressovsky, H., 103–105
Toong, Y.C., 570
Tracey, M.L., 122, 126
Trippa, G., 23
Trivers, R.L., 567
Turelli, M., 218, 494, 498, 499, 501–503, 529, 554, 595
Turner, J.R.G., 211, 212, 570, 574, 576

Upholt, W.B., 332
Uyenoyama, M.K., 416, 556, 560, 563, 566

van Delden, W., 224
van Noordwijk, A.J., 241, 242
Van Valen, L., 515
Via, S., 476
Voelker, R.A., 52, 108
Vrijenhoek, R.C., 603

Wade, M.J., 496, 529, 561, 562, 563
Wahlund, S., 282
Wallace, B., 105, 106, 118, 119, 525
Walsh, J.B., 351, 388, 404
Walton, W.E., 535
Waples, R.S., 322
Watanabe, T., 18, 580, 604
Watt, W.B., 546
Watterson, G.A., 128, 130, 133, 140, 142, 143, 228
Wayne, R.K., 462
Wehrhahn, C.F., 136
Weiler, H., 493
Weinberg, W., 32
Weir, B.S., 51, 55–57, 136, 219, 264, 266, 267, 269, 293, 297, 300, 322
Weiss, M.L., 317, 319
Wellauer, P., 395
Weller, J.I., 506, 507
Werren, J.H., 560
White, M.J.D., 591, 595, 596
Whittam, T.S., 21, 143, 228, 376
Widegren, B., 402
Wijsman, E.A., 309
Wilkinson, G., 498
Williams, S.M., 393, 394
Wilson, A.C., 347, 348, 516, 593, 602
Wilson, D.S., 323, 529
Wilson, E.O., 567
Wilson, S.R., 221
Wilton, A.N., 491
Woese, G.R., 330, 393
Wolfe, K.H., 363
Wolstenholme, G.E.W., 580
Wooten, M.C., 379
Workman, P.L., 269, 277
Worton, R.G., 395
Wright, J.E., 603
Wright, S., 61, 66, 69, 70, 71, 72, 79, 80, 82, 89, 166, 207, 208, 246, 249, 281, 293, 297, 309, 313, 314, 317, 322, 323, 454
Wu, C.-I., 117, 363, 558
Wunsch, C.D., 345
Wynne-Edwards, V.C., 561

Yamaguchi, O., 18, 252, 253
Yamazaki, T., 223
Yanofsky, C., 358
Yokoyama, S., 365
Young, S.S.Y., 191, 493
Yunis, J.J., 598, 599

Zlotogora, J., 324
Zouros, E., 52, 553, 604
Zuckerkandl, E., 360

Subject Index

ABO blood group, 40
Absorbing states, 67
Acinonynx spp. (cheetah), 22–23, 84, 462
Acquired immunodeficiency syndrome (AIDS), 365–366, 372
Adaptation, 2,4
Adaptive evolution, 110, 118, 147
Adaptive suicide, 563
Adaptive topography, 166, 213–215, 495
  shifting balance theory and, 323–326
Adenylate kinase-1 gene (ADK-1), 20, 126
ADH, *see* Alcohol dehydrogenase
Agarose gel electrophoresis, 331–332, 335
Age-structured populations, 35, 530–538
Agrostis tenuis (bent grass), 550–553
Albinism, 99
Albumin, 348
Alcohol dehydrogenase (ADH) 30, 335–340, 372, 377, 382, 509–511
Allele(s)
  additive effects, 120,451
  effective number of, 125
  modifier, 216–217, 490
  multiple, 168–177, 299
  private, 319–320, 322
  rare, 19, 137–138, 320
  selectively neutral, 122–126, 132–134
  X-linked, 43–45, 79
Allele frequency, 17–19
  calculating, 17–19
  clines and, 281, 319, 544–548, 551
  conditional average, 322–323
  constancy under Hardy-Weinberg, 35
  equilibrium, 180–182, 192–193
  estimating using the EM algorithm, 40–42
  genetic distance, 301–306
  genotype frequency and, 18, 186–190
  inbreeding and, 244–250, 268–270
  migration and, 306–323
  mutation pressure and, 110–114
  natural selection and, 151–220
  neutral theory and, 122–130, 136–143
  random genetic drift and, 61–74
Allele generation, 108
Allopatric speciation, 591–595
Allozygosity, 236–237, 253, 257
Allozyme(s)
  definition, 19
  genotypes, estimating fitness of, 150, 538–555
  linkage disequilibrium and, 51–53
  mutation and, 96, 108–109
  polymorphism, 19–24, 136, 150, 161–162, 174, 302
  selection and, 538–555
  self-fertilization and, 257
Altruism and kin selection, 563–568
Alu family of repeats, 402–406
Amanita mushrooms, 549–550
Amino acid(s)
  DNA code, 98
  electrophoresis and, 9–11
  expected proportion of site differences 355–356
  molecular clocks and, 358–367
  patterns of substitution, 368–375
  polymorphism, 9–11, 109
  proportion of site differences (D) 354–356
  replacement rate (lambda), 109, 354–357
  selective constraints on, 228–230, 342–343, 376–377
  sequence, 330
Amplification, 14, 384
Amylase, 18, 52, 185, 223, 338, 507
Anemia
  sickle cell, 7, 99, 162, 169–170, 198, 284
  thalassemia, 339, 391, 403
Aneuploid gametes, 99, 163
Angiosperm evolution, 424–425
Animal husbandry, 582
Ankylosing spondylitis, 54
Antagonistic pleiotropy, 504, 535, 536
Anthoxanthum odoratum (sweet vernal grass), 550–553
Antibiotic resistance, 413, 580
Antigenic specificity, 347
Antiserum, 347–348
α-1 Antitrypsin deficiency, 283–284
Apolipoprotein (human), 507–508
Archaebacteria, 330
Artificial selection
  asymmetric response, 492
  bilateral symmetry, 447–448
  change in mean phenotype, 456–457
  correlated responses, 489–493
  directional selection, 441, 448–450
  gene frequency change and, 454–456
  in genetically uniform populations, 441–442
  identified genes and, 507–511
  individual, 442–446
  intensity, 483
  limits to, 261, 446–450
  linearity of response, 479
  multivariate characters, 493–496
  prediction equation, 457, 442–446

relaxed, 446–447
response, 443–446, 456–457
selection differential, 457
selection index, 489
truncation selection, 442–446
Ascomycetes (fungi), 101
Ashkenazi Jews, 284, 324–325
Assortative mating, 195, 271–274
negative assortative mating (disassortative mating), 271, 274–277
Autoradiography, 15, 28, 331
Autosome, 45
Autozygosity, 74–82
inbreeding and, 238–244
infinite alleles model and, 123, 127
probability of, 81–82, 89, 248–249
Avena barbata (oats), 260

Baboon, 417–419
Backcrossing, 255, 258
Background genes, 223–224, 272
Bacteria
antibiotic resistance, 580–586
selection in, 111–114, 183–184, 411, 541–542
transposable elements in, 409–413
Bacteriophage
coevolution with bacteria, 586–587
λ, 585
M13, 341
T5, 112, 183
Balancer chromosomes, 104–105, 118, 121, 157–158
Balancing selection, 381–382
Barley (*Hordeum vulgare*), 260
Base substitution, *see* Nucleotide substitution
Batesian mimicry, 570–576
Bennett-Binet effect, 267
Biased gene conversion, 101, 405
Bilateral symmetry, 447–448, 474
Binomial distribution, 62, 66–67, 88–89, 207–208
Biological constraints, *see* Constraints to evolution
Biological species concept, 588
Biometricians vs. Mendelians, 5
Biston betularia (peppered moth), 577–580
Block plot, 347, 394
Blood group polymorphisms
ABO, 40
Duffy, 308–309
Kell, 308
Kidd, 308
MN, 37–38, 308–309
Rh, 38–39, 186
Bottlenecks, 83–84, 86, 364, 547
Branch point effects, 542
Brassica, 275, 424
Breeding programs, 432–433, 435
see also Artificial selection
Bristles, of *Drosophila*
length, 393, 394
number, 431–434, 445–447, 488
Broad-sense heritability, 464–465, 470–471

Caenorhabditis elegans (nematode), 27
Caffeine, 112–113
Candidate genes, 507
Carp, 359, 360
Carrying capacity, 518, 519, 595
Cat, 419
Central limit theorem, 433–435
Chain-termination DNA sequencing method, 14–15, 341–342
Chaos, 522
Charge state model, 135–136
Cheetah (*Acinonynx* spp.), 22–23, 84, 462
Chemostat, 111–114, 183–184, 411, 541
Chicken, 359, 360, 470, 489
Chimpanzee, 347, 367, 598–599
Chi-square test, 37, 115–116, 135, 140, 281
Chloroplast DNA, 363, 424–425
Cholesterol and apolipoprotein, 507–508
Chromosome
balancer, 104–105, 118, 121, 157–158
in mice, 192–193
multigene families and, 384–395
repetitive DNA sequences and, 399–406
X, 43–45, 103–105, 243–244, 393–395, 408, 557–559, 600–601
Y, 92, 393–395, 600–601
Chromosome abnormalities, 96, 99, 163
Chromosome replacement lines, 491, 537
cis (coupling phase), 209
Cleavage sites, *see* Restriction site
Clines in allele frequency, 281, 317–319
Coadapted gene complex, 99
Coalescent, 69, 90–92, 130–131, 134–135
Coancestry, coefficient of, 238–239, 258–259, 469
Cod (*Gadus morhus*), 285–287
Codon preferences (codon bias), 371
Coefficient
of coancestry, 238–239, 258–259, 469
correlation, 53–54, 437
of inbreeding, 80, 236–261, 295–297, 564
of linkage disequilibrium, 47–57, 211–216
regression, 331, 362, 435–438, 468
of relationship, 564
selection, 149, 155–156, 199
of sexual selection, 196–197
Coevolution, 569–576
bacteria and bacteriophage, 586–587
Colias (butterfly), 547
Color blindness, 45, 391–392
Computer algorithms
chi-square (χ^2), 38
EM, 40–42
hashing, 347
maximum likelihood, 478
Needleman-Wunsch, 345–347
Stewart, 140–143
Concerted evolution, 101, 384–388, 393–395, 406
Concordance, of traits in twins, 484–485
Conditional average allele frequency, 322–323
Confidence interval, 134, 140–141, 143
Consensus sequence, 336, 340, 402
Constraints to evolution
ecological, 2, 514–517
genetic, 228–230, 342–343, 374, 376–377, 474–476
Continuous traits, 7, 430–435
Controlling elements in maize, 407
Conversion (gene), 384–393, 405
Cooperation, evolution of, 568–569
Coot clam (*Mulinia lateralis*), 553–554
Copepod, 535, 544, 547
copia element, 413–415
Correlated response to selection, 489–493, 537

Correlation coefficient
defined, 53–54, 437
of heterozygosity under drift, 132
linkage disequilibrium and, 53–54
of uniting gametes, 249
Corvini (crows and ravens), 329
Courtship feeding, 195
Covariance, 289–293, 436–437
between relatives (quantitative traits), 466–472
cpDNA, *see* Chloroplast DNA
Cramer's rule for linear equations, 172
Crickets (*Gryllus*), 597
Crossing over, 101
double, 53
unequal, 384–393
Cryptic genes, 395
Crystallization (genome), 218
Cumulative selection differential, 443
Curly (*Cy*) wing mutation, 105–106, 118–119, 250–252
Cystic fibrosis, 39, 284
Cytochrome, 374, 404
Cytotype, and hybrid dysgenesis, 415–416

Daphnia, 500, 530
Darwin's finches, 497, 500
Deer mouse (*Peromyscus maniculatus*), 90, 546
De Finetti diagram, 35–36, 46–47, 161, 172, 247, 276–277, 283
Degeneracy of nucleotide sites, 368, 370
Degrees of freedom, 38,39
Deme, *see* Subpopulation
Demographic theory, 516, 517, 530–533
Density dependent natural selection, 517–524
Diabetes mellitus (IDDM), 228, 432, 481
Diallel design, 538
Diallelic genes, 46–49, 56–57
Dideoxynucleotides, 342
Differential mating success, 194–198
Diffusion equation, 71–74, 312–314, 323, 351
Digenic measures of descent, 263–268, 293
Dimer, 11, 21
Diploidization, 603
Directional selection, 166, 202, 208, 441, 448–450
Dirichlet distribution, 376
Disassortative mating (negative assortative mating), 271, 274–277
Dispersed repetitive DNA, 402–406
human *Alu* family, 402, 406
rabbit C repeats, 402
SINEs, 402
Distance matrix, 379
DNA
amplification, 343–344
chloroplast (cpDNA), *see* chloroplast DNA
dispersed repetitive sequences, 402–406
electrophoresis and, 9, 11–12, 331–332
fingerprinting, 27
fragment size and migration distance, 331–332
heteroduplex, 101,345
localized repetitive sequences, 399–402
melting temperature, 344
minisatelliates, 26–27
mitochondrial (mtDNA), *see* Mitochondrial DNA
multigene families, 384–395
nucleotide substitution patterns, 368–375
polymerase chain reaction, 343–344
primers, 344
probe, 12, 331
repair, 101–103, 109
restriction fragment length polymorphism (RFLP), 13, 332–339
satellite, 399–402
selfish, 399, 403, 406
transposable elements, 407–420
DNA-DNA hybridization, 344–345
DNA sequence and sequencing
alignment methods, 345–347
databases, 14, 339
divergence, 334–335
identity blocks, 347
inference of recombination, 382–383
inference of selection, 376–377
Maxam-Gilbert method, 14–16
phylogenetic inference and, 329–330, 344–349, 366–367, 377–382
Sanger method (dideoxy chain termination), 14–15, 341–342
variation, 11, 14, 332–339
Dog, 359, 360
Dominance
Hardy-Weinberg principle and, 38–39
molecular basis of, 539–542
mutation selection balance and, 199
quantitative traits and, 451, 463–466
recessive lethals and, 199–200
viability polygenes, 121
Dominance variance, 463–464
Dot plot, 346
Double heterozygote, 209, 260, 264
Double homozygote, 267–268
Down syndrome, 6
Drosophila
allozyme polymorphism in, 19–21
antagonistic pleiotropy, 537–538
artificial selection, 445–447, 462–463
balancer chromosomes and, 104–105
enzyme activity and, 7, 490–491, 542
fitness-related traits, 184–186
mutation-selection balance and, 104–107
repetitive DNA sequences, 384, 393–395, 400–401
Drosophila arizonensis, 604
Drosophila ezoana, 400
Drosophila heteroneura, 138
Drosophila littoralis, 400
Drosophila mauritiana, 112, 343, 600–601
Drosophila melanogaster
alcohol dehydrogenase, 30, 335–340, 372, 377, 382, 509–511
amylase, 18, 52, 185, 223, 338, 507
balancer chromosomes and, 105–107, 118–122
bar eye mutation, 104
bobbed, 393, 394
bristle length, 393
bristle number realized heritability, 445–447, 488
bristle phenotype, 431–434
chromosome abnormalities, 157–158, 163–165
copia element, 413–415
Curly (*Cy*) wing mutation, 105–106, 118–119, 250–252

density dependent selection, 522–524
ebony and mating isolation, 590
enzyme activity variation, 490–491, 542
esterase-6, 548
GATA repeat, 401
glucose-6-phosphate dehydrogenase (G6PD), 224, 491, 507
α glycerophosphate dehydrogenase, 542
hybrid dysgenesis, 415–416
inbreeding, 250–252
linkage disequilibrium and, 52–53, 55, 450
magnification of rDNA, 394
meiotic drive and, 191, 193
migration rate, 305, 320–321
Notch, 338
P elements, 413–416
phenotype abnormalities, 8
6-phosphogluconate dehydrogenase (6PGD), 491, 509
polymorphism, 19–21
random genetic drift and, 64–66
rates of mutation in, 108
rDNA, 384, 393–395
recessive lethals in, 199–200
restriction site variation, 25–26, 335–339
roo, 408, 409
satellite DNA, 400
sexual selection and, 198, 498
technical knockout, 339
thorax length inheritance, 462–463
transposable elements and, 27, 413–416
white, 337–338, 409
wing length and selection, 498
wing vein selection, 448–449
Y chromosome 394, 600–601
Drosophila mercatorum, 393
abnormal abdomen, 547
Drosophila mojavensis, 604
Drosophila nasutoides, 400
Drosophila nebulosa, 414
Drosophila orena, 343
Drosophila persimilis, 42, 589
Drosophila pseudoobscura
alcohol dehydrogenase, 336–337
bristle number inheritance, 472–473
fitness, 122
inversion polymorphism, 99, 175, 517
migration rate, 305
norm of reaction, 472–473
reproductive isolation, 589–590
sequential electrophoresis, 24, 141
temporal variation, 530
xanthine dehydrogenase (XDH), 141–143
Drosophila putrida, 550
Drosophila recens, 550
Drosophila sechellia, 377, 600–601
Drosophila simulans, 343, 595, 600–601, 604
Drosophila tripunctata, 549–550
Drosophila willistoni, 126, 322, 414
Duffy blood group, 308–309
Duplication, 96, 386–393
Dynamics of population growth, 518–519

E. coli, see Escherichia coli
Echidna, 359, 360
Ecological opportunity, 515
Ecological time, 517
Eelpout (*Zoarces viviparus*), 138, 224–227
Effective number of alleles, 125
Effective number of genes affecting quantitative trait, 485–489
Effective population number
and allele fixation, 64, 82, 123
drift in phenotypic characters and, 498–501
fecundity and, 88–89
fluctuating population size, 83–85
migration and, 301
natural selection and, 89
neighborhood size, 90
overlapping generations, 87–88
selection limits and, 447
sex ratio and, 85–87
Eigenvalue, 68, 160, 179, 202–206
Eigenvalue effective size, 82–83
Electrophoresis, 9–12, 136, 331–332, 335
allozyme variation and, 19
in DNA sequencing, 15
sequential, 23–24, 141
two-dimensional technique, 24
Electrophoretic mobility, 10, 18, 136, 141
Elephant seal, 84
El Niño, 497, 500
EM algorithm, 40–42
EMBL, 14
EMS mutagen, 121
Environment
heritability and, 460, 472–473
interaction with genotype, 460–461
Environmental covariance, 439, 490
Enzyme(s)
activity, 7
activity variation, 490–491, 542
dimers, 21
electrophoresis and, 10, 21
Group I vs. Group II, 24
kinetics and metabolic control, 539–542
monomeric, 11,21
multimeric, 11
restriction, 11
thermolabile, 23
Enzyme kinetics and control theory, 539–542
Epistasis, 209–216, 465
Epistatic selection, 132
Epistatic variance, 465
Equilibrium
autozygosity, 124–125
Hardy-Weinberg, 31–40
interior, 169, 178–180, 189–190, 192, 196–199, 202–206
natural selection and, 154–164, 184–186
polymorphic, 154, 159, 169
self-incompatibility and, 275–277
stability determination, 159–161
Equilibrium autozygosity, 124–125
Error, type I and type II, 138–139
Escherichia coli
allozyme variation in, 19, 21, 183–184
bgl and β-glucoside metabolism, 395
evolution of tetracycline resistance, 584–585
β-galactosidase and fitness, 541–542
gnd, 376
IS elements, 409–413
permease, 541
resistance transfer factor, 580–586
SOS DNA repair, 101–103

transposable elements in, 409–413
trp A, 358, 370
uvr and DNA repair, 101–103
Esterase, 24, 108, 224–227, 260, 547, 548
Estimator(s), 236, 281, 437
Euglena gracilis, 339
Euler equation, 533, 534
Euphydryas editha (butterfly), 89, 530
European Molecular Biology Laboratory, 14
Evolution
concerted, 101, 384–388, 393–395, 406
defined, 1–2
mtDNA and cpDNA, 421–426
repetitive DNA sequences, 384–395
see also Amino acid substitution; Nucleotide substitution; Selection
Evolutionarily stable strategy, 569
Evolutionary species, 588
Evolutionary time, 517
Ewens sampling formula, 126, 139–140, 141, 143
Ewens-Watterson test, 130, 140–143, 376
Exon, 340, 342
Exon shuffling, 103, 397

F statistics, Wright's
analysis of variance, 298–300
estimating, 297–298
Ewens-Watterson test and, 140–143
partitioning of, 298–299
variance-covariance matrix and, 299–300
Fecundity, 149
additive, 187
mean, 190
multiplicative, 188, 190
parameter, 187
symmetric, 189–190
see also Natural selection
Fecundity curve, 531
Fibrinogen, 374
Fisher-Wright model, *see* Wright-Fisher model
Fitness
additive, 187, 199, 212–216
components of, 148–149
estimation of, 149–150
inclusive, 564–565
induced, 214
marginal, 165, 168–170, 175, 211
mean, 119–121, 164–168, 173–176, 206–210, 213–214
molecular variation and, 228–230
multilocus, 217–219
multiplicative, 215
mutation rate and, 110
as quantitative trait, 121
relative, 154, 164, 175
spontaneous mutation and, 118–122
zoo and wildlife management and, 89
Fixation, 62–65, 69, 72–74, 82, 192–193, 206–207
Fixation index, 123, 236, 284, 288, 294, 297, 302
Flour beetle (*Tribolium castaneum*), 435–436, 447, 488, 562
Fluctuating asymmetry, 462
Fluctuating neutral space, 364
Fluctuating selection, 364, 529–530
Flux and metabolic control, 540–541
Founder effects, 84, 284, 547
Founder-flush cycles and speciation, 590, 595
Four-cutters, 11–12
Frameshift mutation, 99
Frogs and toads, 347
Fundamental theorem of natural selection, 164–168, 173, 176–177, 182, 190, 206
Fundulus heteroclitus, 544

Gainj tribe, 301–302
β-Galactosidase, 541–542
Galapagos Islands, 497
Gametic frequency, 48–49, 51, 292–293
Gametic selection, 149, 190–194
Gametophytic
selection, 204–206
self-incompatibility, 274–275
Gap penalty, 345
Gaucher disease, 324–325
Gaussian distribution, 433–435
Gegenbauer polynomials, 72
Gel electrophoresis, *see* Electrophoresis
Gelada, 419
GENBANK, 14
Gene(s)
autosome, 45
counting, 41,55
diallelic, 46, 48–49
effective number, 125
fixation, 62–65
frequency, 18
modifier, 110, 216–217
nontranscribed, 114
regulatory, 24
substitution, 96–97; *see also* Amino acid substitutions; Nucleotide substitutions
target, 110
see also Allele(s)
Gene amplification, 384
Gene conversion, 101–102, 384, 388, 389, 390, 405
Gene counting, 41,55
Gene duplication, 96, 386–393
estimating time of, 390
Gene expression, 97
Gene flow, 551
speciation and, 590–597
see also Migration
Gene genealogy, 377, 379, 382
Gene identity, 142
Gene pool, 588
Genealogy, 90–92, 130–131, 134, 336
Genetic
architecture, 474
coadaptation, 52
distance, 301–305
divergence, 301, 306, 312, 317, 324
diversity, 126
Genetic code, 2, 97–98
table, 98
Genetic covariance, 469, 490
genetic mechanisms of, 492–493
Genetic disorders, 99
see also under specific disorder
Genetic drift, *see* Random genetic drift
Genetic variance components, estimation of, 476–480
Genetical species, 588
Genic variance, 80
Genotype-environment interaction, 460–461
Genotype frequency
biased gene conversion and, 101
defined, 32
De Finetti diagram and, 35–36
inbreeding and, 244–249, 261–263, 268–271
mating patterns and, 194–198
mixed mating model and, 261–263
natural selection and, 148, 184–186

random genetic drift and, 81–82
see also Hardy-Weinberg principle
Globin
α, 359, 360, 362
β, 7, 169, 173, 198, 361
deer mouse and altitude adaptation, 546
Glucose-6–phosphate dehydrogenase (G6PD), 224, 491, 507
deficiency, 99
Glutamate oxaloacetate transaminase, 547–548
Goldenrod, 488
Goodness-of-fit, *see* Chi-square test
Gorilla, 347, 367, 598–599
Grasshoppers, flightless (*Morabinae*), 596
Great tits, 241–242
Group selection, 529, 560–563
Guinea pig, 362, 454, 460
Guppies, 537
Gynogenesis, 603

Habitat selection, 528
Haldane-Muller principle, 201
Half-sib mating, 258
Hamilton's rule, 565–566
Haploid natural selection, 182–183
Haplotype, 26, 336, 338
Haplotype diversity, 336
Haplotype partitioning, 382
Hard selection, 525–527
Hardy-Weinberg principle, 31–40
assumptions of, 31–32, 36, 64
gene linkage and, 45–55
hypothesis testing and, 61
inbreeding and, 244–250
multiple alleles and, 39–43
mutation and, 97
natural selection and, 151
random genetic drift and, 70
self-fertilization and, 244–249
Wahlund's principle and, 281–293
X-linked genes and, 43–45
Harmonic mean, 83, 179, 182
Harmonic mean fitness, 528
Hashing, 347
Hawaiian *Drosophila*, 401, 488, 516, 593, 594, 600
Heavy metal tolerance, 548–553
Hemoglobin, 162, 169, 357, 369
polymorphism in cod, 285–286
see also Globin; Sickle-cell anemia
Hemoglobinopathies, 339
Hemophilia, 403
Heritability
broad-sense, 464–465, 470–471
concordance in twins and, 484–485
effect of dominance on, 452–453
effect of environment on, 460, 472–473
effect of gene frequency on, 465–467, 479
estimation of, 476–480
examples in domestic animals and plants, 470
familial resemblance and, 435–441
fitness, 470
gene-environment association, 461
genetic basis, 440
human traits, 471
maximum likelihood methods, 477–478
Mendelian traits and, 465
narrow-sense, 439–441, 464–468, 470
parent-offspring regression, 435–441, 468
realized, 445–446
twin studies, 480–481
Heterodimer, 11
Heterokaryotype, 42
Heteroplasmy, 425–426
Heterosis (hybrid vigor), 261
Heterostyly, 276
Heterozygosity
advantage, 173–174, 181, 198, 202, 213
allozyme polymorphism, 19–22
assortative mating and, 274
average, 143
divergence among subpopulations, 302
double, 53, 209
growth rate and, 553–555
inbreeding and, 74–75
inbreeding effective number and, 77–82
inferiority, 163–164, 166
infinite alleles model and, 122–126
negative assortative mating and, 277
neutral mutations and, 130–132, 137
neutral theory and, 136–143
of nucleotides, 25
partial selfing and, 262
steady state with mutation and drift, 122–126
virtual, 124, 336
X-linked genes and, 79
Histone, 357, 369
HLA (Human Leukocyte Antigen), 54, 219, 228, 319
Homeotic mutants, 6
Homodimer, 11
Homokaryotype, 42
Homozygosity
in allele frequencies, 124–126
balancer chromosomes and, 118
inbreeding and, 235–250, 267–268
infinite alleles model and, 136
neutral theory and, 136–143
Wahlund's principle and, 282
Hordeum vulgare (barley), 260
Horses
coefficient of inbreeding, 242–243
teeth, 501
House-of-cards model, 503
Human
ABO blood group, 40
Alu and *Kpn* repeats, 346, 402–406
apolipoprotein and cholesterol, 507–508
birth weight, 496–497
broad sense heritabilities, 471, 480–481
color blindness, 45, 391–392
cystic fibrosis, 39, 284
diabetes, 228, 432, 481
Duffy blood group, 308–309
Gaucher disease, 324–325
α globin, 359, 360, 388, 389, 391, 392
β globin, 391
Huntington disease, 530
inborn errors of metabolism, 543
karyotype, 598–599
lysosomal storage disorders, 324–325
MN blood group, 37–38, 308–309
mucolipidosis, 324–325
phenylketonuria (PKU), 543
phosphoglycerokinase (PGK), 404
pyloric stenosis, 483
Rh blood group, 38–39, 186
schizophrenia, 432
sickle-cell anemia, 7, 99, 162, 169–170, 198, 284
skin color, 488

Tay-Sachs disease, 284, 324–325
thalassemia, 339, 391, 403
twin studies, 480–481
xeroderma pigmentosum and DNA repair, 101
Human immunodeficiency virus (HIV), 365–366, 372
Human leukocyte antigen (HLA), 54, 219, 228, 319
Huntington disease, 199–200, 530
Hybrid corn, 261
Hybrid dysgenesis, 415–416
Hybrid sterility, 600–601
Hybrid vigor (heterosis), 261
Hybrid zones, 597
Hybridogenesis, 603
Hyla, 348
Hymenoptera and kin selection, 566–567
Hypergeometric function, 72

Identity block, 347
Identity by descent, 75
in multigene families, 386, 387
see also Autozygosity
Identity disequilibrium, 264–268
Immunological distance, 347–349
Inbreeding
autozygosity, 75
coefficient, 80, 238–244
depression, 253–255, 554
effects of, 250–255
evolutionary effects, 235
genetic variation and, 7
genotype frequencies and, 244–245, 261–263
random genetic drift and, 74–82
systems of, 255–263
viability in *Drosophila*, 252
see also Autozygosity
Inbreeding coefficient, 80–81, 235–261, 564
calculation from pedigree, 238–244
increase in, 262, 257
kin selection and, 565–566
Wright's F statistic and, 295–297
for X-linked genes, 243–244
see also Fixation index
Inbreeding depression, 253, 554
Inbreeding effective size, 82–89
Inclusive fitness, 564–565
Industrial melanism, 577–580
Inebriometer, 509–511
Infinite alleles model, 122–132, 136–143, 352, 376, 380
Infinite sites model, 132–135
Influenza virus, 364–365
Inheritance, polygenic, 430–435
see also Heritability
Insertion sequences, 409–413
*IS*50, 409
population genetics, theory of, 410–411
Insulin, 330, 374, 375
Intensity of artificial selection, 493
Interdeme selection, 323–326
Interferon, 354, 356, 369
Intrinsic rate of increase, 518, 533
Introns, 17, 342
IQ, inbreeding and, 254
IS elements, 409–413
Island model, 306–316, 320
Isogenic, 106
Isolate breaking (population fusion), 283–284
Isolation by distance, 305, 317–319

Jacobian matrix, 178–179
Jensen inequality, 294–295
Just-so stories, 538–539
Juvenile hormone, 570

K-selection, 523, 537
Kalam tribe, 301–302
Kangaroo, 359, 360, 399
Karyotype, 42, 597, 599
Katydids, 195
Kimura-Lande-Fleming model, 502–503
Kin selection, 529, 563–568
Klenow fragment, 341
Kpn sequences, 402–406
Krebs cycle, 538

Leghemoglobin, 360
Lethal hybrid rescue, 604
Leucine amino peptidase-5 gene (LAP-5), 126
Levene model (soft selection), 525–528
Liability and threshold traits, 481–485
Life cycle, 148–149
Life history, 535–538
Linear pressures, 314
LINEs
human *Kpn* family, 403
L1 repeats, 403, 406
and mutation, 403
Linkage disequilibrium, 47–57, 336, 339
coefficient of, 211
estimating, 55–57
founding population and, 54
HLA complex and, 54
infinite alleles model and, 130–132
mixed mating model and, 266–268
MN blood group and, 54–55
multilocus, 219–220
paracentric inversions and, 51
parameter (*D*), 48–50, 53, 57, 260–261
Pearson correlation coefficient and, 53–54
response to selection and, 450
Rh blood group and, 54–55
self-fertilization and, 260
sexual selection and, 194–195
Wahlund effect and, 291–293
Linkage equilibrium, 46–52, 260
Linkage phase, 209
Lizards, 426
Local stability, 159–161, 522
Localized repetitive DNA, 399–402
Logistic equation, 518, 519
Lolium perenne (ryegrass), 552
Lucilia cuprina (sheep blowfly), 474
Lycopersicon (tomato), 393,488, 506
Lynx, 84–85
Lysosomal storage disorders, 324–325
Lytic cycle, 586

Macaque, Japanese, 137–138
Macroenvironment, 459–460
Magnification (rDNA), 394
Maize, 115, 424, 470, 488
oil content selection, 442
Major histocompatibility complex (MHC), 23, 186, 192, 377, 397
Malaria, 162, 169–170, 198, 199
Malic enzyme, 547–548
Malthusian parameter, 514, 517
Mandrill, 419
Mapping quantitative traits, 504–507
Markov chain,
coalescence, 69, 91–92
Wright-Fisher model, 67–70
Mass action, 275–276
Maternal-fetal incompatibility, 186

Mating, and sexual selection, 194–198
Maxam-Gilbert sequencing, 14–16
Maximum likelihood estimates 335, 376
  EM algorithm, 40–42
  gene counting, 55
  genetic variance components, 477–478
  linkage disequilibrium, 56
Mean, 436
Meiotic drive, 149
  viability selection and, 201–204
  gametic selection and, 190–194
Mendelian inheritance, 35, 435
Mendel's law of segregation, 148, 151, 191, 202, 204, 239–240
Meristic traits, 431, 435
Metabolic control theory, 539–542
MHC, *see* Major histocompatibility complex
Michaelis constant, 540
Microcomplement fixation, 347–349
Microenvironment (and quantitative traits), 459–460
Midparent, 439, 468
Migration, 281–326
  asymmetric, 301
  estimating rates of, 319–323
  infinite alleles model and, 124
  one way, 308–310, 316
  rates of, 305, 307, 311, 316
Mimicry, 570–576
  and modifier genes, 574–576
Minisatellites, 26–27
Mismatch distribution, 133–135
Missense mutation (nonsynonymous substitution), 97–98
Mitochondrial DNA (mtDNA), 29, 347, 421–426
  coalescence and, 92
  Eve, 92
  genetic code, 422
  heteroplasmy, 425, 426
  house mouse (*Mus musculus*), 29
  human, 423, 424
  paternal transmission, 426
  phylogenetic studies, 29
  population genetics theory of, 425–426
  primate phylogeny and, 366–367
  sequence divergence in, 29
  substitution rate, 422–423
Mixed mating model, 261–264
MN blood group, 37–38, 308–309
Modifier gene, 110, 490
Molecular clocks, 358–367
  calibrated by DNA-DNA hybridization, 418–419
  episodic, 364
  generation time effect, 363
  and the Poisson process, 362–364
  variation among genes, 362
  variation among lineages, 363
Molecular drive, 400
Monarch butterfly, 572–574
Monomeric enzymes, 11,21
Monomorphism, 18, 132
Moran model, 87–88, 376, 377–378
Mormon crickets, 195
Morphological evolution, 430–431, 515
Morphological variation, 6
Moths and industrial melanism, 577–580
Mouse, 354, 356
  allozyme polymorphism, 19, 24, 29
  body weight inheritance, 474–475, 479
  coat color mutations, 107–108
  meiotic drive and, 191–193, 201
  t-allele, 191–193, 201
  translocations, 100–101
mtDNA, *see* Mitochondrial DNA
Mucolipidosis, 324–325
Müllerian mimicry, 570
Multifactorial traits, *see* Quantitative traits
Multigene family, 101, 347, 384–395
  population genetic models, 386–388
  and molecular drive, 400
  and selfish DNA, 399, 406
Multigene superfamilies, 397–399
Multimeric enzymes, 11
Multiple alleles, 168–177
  fundamental theorem and, 176–177
  polymorphism, 174
  Wahlund's principle and, 288–292
Multiple substitutions, 356
Multivariate Gaussian distribution, 493–494
Multivariate selection theory, 493–496
Mus domesticus (mouse), 19, 24, 29, 192–193
Mus musculus (mouse), 10, 192–193
Mussel (*Mytilus edulis*), 544–545
Mutation, 96–103, 357
  allele frequencies and, 97, 135–136
  balance with random genetic drift, 122–130
  balance with selection, 198–201
  beneficial, 110
  bidirectional, 114, 316
  classes of, 97–103
  defined, 97
  –drift equilibrium, 499
  fitness and, 118–122
  frameshift, 99
  island model and, 314–316
  missense (nonsynonymous mutation), 97, 103
  mobility variant, 108–109
  neutral, 128, 142
  nonrandomness and, 114–117
  nonsense, 99, 103
  null, 108–109
  radiation induced, 105
  recurrent, 97, 199, 201
  reversible, 113–114
  –selection balance, 199–201, 501–504
  selectively neutral, 97, 123
  silent (synonymous), 97, 103
  spontaneous, 118–122
  transition, 97, 116, 118
  transposition, 96
  transversion, 116
  see also Amino acid substitution; Nucleotide substitution; Infinite alleles model
Mutation load, 201
Mutation pressure, 97, 110–114, 122–123
Mutation rate, 97, 103, 123
  adaptive, 109
  of allozyme genes, 108
  estimation, 103–110
  genetic variation in, 110, 139
  mutation pressure and, 110–114
  neutral alleles, 139
  spontaneous, 97, 103–110
  of X-linked lethal recessives, 103–107
Mycophagous *Drosophila*, 549–550
Mytilus edulis (mussel), 544–545

Narrow-sense heritability, 439–441, 464–468, 470
Natural selection
age-structured populations 533–534
allozymes 543–548
coefficient, 149, 155–156, 199
defined, 148–149
density dependent 517–524
directional, 154, 156, 166, 202–203, 205–209
fecundity, 149, 186–190, 196–197
frequency dependent, 184–186
fundamental theorem, 164–168, 173, 176–177, 182, 190, 206
gametic, 149
in haploids, 182–183
hard selection 525–527
inference from phenotypic data, 496–498
interdeme, 323–326
K-selection 523
nucleotide sequence data 376–377
overdominance, 158–162
purifying 380, 381
r-selection 523
sexual selection, 194–198
shifting balance theory and, 323–326
soft selection 525–528
sperm displacement, 193
sporophytic and gametophytic stages
stabilizing 380, 381, 501, 504
temporal variation 529–530
underdominance, 163–164
viability, 151–186
Natural selection, opposing forces, 198–209
haploid and diploid, 190–191, 204–206
mutation-selection, 199–201
natural selection and inbreeding, 268–271
selection and migration, 309–311
viability and fecundity, 206
viability and meiotic drive, 201–204
viability and migration, 309–311
viability and random genetic drift, 206–209
Needleman-Wunsch algorithm, 345–347
Negative assortative mating (disassortative mating), 271, 274–277
Neighborhood size, 90
Nematode (*Caenorhabditis elegans*), 27
Neomycin resistance, 409
Neurospora (fungi), 101
Neutral alleles, 122–123
Neutral mutation, 137, 150
Neutral theory of molecular evolution, 136–143, 349–353
average time to fixation, 351
average time to loss of an allele, 351–352
coalescent properties, 90–92, 134–135
equilibrium homozygosity, 352
fixation probability, 350
infinite alleles model and, 122–126, 136–143
infinite sites model, 132–133
principles, 349–353
substitution rate, 350–351
tests of, 136–140
two neutral loci, 136–143
Neutrality hypothesis, 136–143
Newt, 359, 360
Nick translation, 12
Niemann-Pick syndrome, 324–325
Nitrocellulose filters, 331, 336
Nonhomologous exchange, 395
Nonsense mutation, 99, 103
Norm of reaction, 460, 472–476
Normal density function, *see* Normal distribution
Normal distribution, 433–435
artificial selection and, 452
Normalization, 151
Normalized identity, 303–305
Nova Scotia (*NS*) inversion, 18, 52
Nucleotide(s)
heterozygosity, 25–26, 332–334
mutation and, 96, 109
polymorphism, 25–26, 30, 132–133, 332–334, 342
Nucleotide sequence (*see* DNA sequence)
Nucleotide substitution, 96–97, 334
estimating the rate of, 357–357, 370
expected proportion of site differences, 358
nonsynonymous (missense), 97, 103, 368–370, 377
rate of, 357–358
selective constraints 371, 375, 376
synonymous (silent), 97, 103, 368–370, 377
transition, 97
transversions, 97
Null hypothesis, 38, 61, 116, 138–139

Oats (*Avena barbata*), 260
Oncogenes, 403, 420
Orangutan, 598–599
Orthopterans, sexual selection and, 195
Outcrossing, 250–261, 301
Overdominance, 154, 158, 161–162, 167, 173–176, 179–180, 202, 205
Overdominance in carrying capacity, 520
Overlapping generations, 82, 87–88

P element, 413–416
Panglossian paradigm, 517
Panmictic index, 256
Papilio dardanus (butterfly), 574–576
Paracentric inversions, 18, 42, 51, 162
Parameters, 3, 236, 437
Parapatric speciation, 595–596
Parental expenditure and sex ratio, 555
Parent–offspring regression, 435–441, 468
Partial selfing, 261–264, 266
Paternal transmission, 426
Paternity analysis, 27
Pearson correlation coefficient, *see* Correlation coefficient
Pedigrees, 238–243, 258–259
Penicillin, 582
Peromyscus maniculatus (deer mouse), 90, 546
Phaseolus (Princess bean) 441, 444
Phenotype(s)
adaptation, 2, 4
assortative mating and, 272
Drosophila bristle number, 431–434
environment and, 430
mean, 433
parent-offspring resemblance, 430, 435–441
variance, 433
Phenotypic plasticity, 472–476
Phenotypic variation, 6, 430–435
Phenylketonuria (PKU), 99, 543
Phlox cuspidata, 263, 296–298
Phosphoglucomutase-2 (PGM-2), 263, 296

6–Phosphogluconate dehydrogenase, 376, 490–491, 509
Phylogenetic inference, 329–330, 344–349, 366–367, 377–382
Phylogenetic trees, 15, 92
  average distance 378
  expected distance 378
  maximum parsimony 379
  mtDNA 424
  primates, 366–367
  realized distance, 378
  UPGMA, 378, 380
Physiological genetics, 538–555
Pig, 470
Pili, 580, 582
Pin, 276
Plants
  assortative mating and, 272, 274–275
  meiotic drive and, 193
  natural selection and, 204–206
  pollen tube competition, 193
  ribulose-1,5–bisphosphate carboxylase, 392–393
  self-fertilization in, 257–263
  self-incompatibility in, 194, 274–277
Plasmids, 413, 580
Pleiotropy, 149, 192, 195, 198
Plethodon dorsalis (salamander), 322
PM hybrid dysgenesis, in *D. melanogaster*, 415–416
Pocket gopher, 423
Poeciliopsis, 603
Poisson distribution, 88–89, 115, 119
Poisson process, 351, 363
Pol gene, 365
Pollen elimination, 277
Pollen tube competition, 193
Polyacrylamide gel, 337
Polygenic traits, *see* Quantitative traits
Polymerase chain reaction, 343–344
Polymorphism(s)
  in allele frequency, 18
  allozyme, 19–22, 136
  complete, 169, 172–174
  estimates of, 23
  meiotic drive and, 201–204
  neutral theory and, 92, 136–142
  nucleotide, 132–133
  overdominance and, 161–162, 180–182
  protected, 178–179
  restriction sites, 25–26, 330–339
  stable, 202–203, 208
Polypeptides, mutation rate in humans, 108–109
Polyploidy, 96
Population(s)
  bottleneck, 83–84, 86
  density, 300–301
  dispersion, 90
  fitness, 323
  fusion (isolate breaking), 283–284
  mean, 434
  pooled, 282, 288–291, 294
  size, 130–131, 301–302
  structure, 281, 300–327
  subdivision, 293–295, 320–321
  see also Effective population number; Subpopulation
Population heritability, 562
Postmating isolation, 589
Power (of statistical test), 55
Prediction equation for artificial selection, 442–446
Premating isolation, 589
Primroses (*Primula*), 286
Princess bean (*Phaseolus*), 441–444
Principle of competitive exclusion, 589
Prisoner's dilemma, 568–569
Probability, 61–62, 67, 69, 134
Probe DNA, 12, 331
Progeny testing, 55
Protein(s)
  amino acid replacements, 354–357
  electrophoresis and, 9
  evolution of, 19–24
  folding, 103
  immunological distance, 347–349
  molecular clock, 358–367
  ribosomal, 24
  see also Enzyme; Hemoglobin; specific protein
Protein conformation, 374
Pseudogenes, 114–115, 395
  processed 404
Pseudolinkage, 603
Punnett square, 32–34, 39, 246
Purifying selection, 142, 380
Purines, 97–98
Pyloric stenosis, 483
Pyrimidines, 97

Quantitative genetics, 5, 121, 431–435
Quantitative traits, 5, 430–435
  additive effects, 451
  additive variance, 463–466
  components of phenotypic variance, 458–465
  covariance between relatives, 466–472
  dominance effects, 451
  dominance variance, 463–466
  environmental variance, 461
  enzyme activities, 490–491, 542
  epistatic variance, 465
  estimation of variance components, 476–480
  genetic basis of, 450–454
  genetic correlation, 490–492
  genotype-environment interaction, 460–461
  genotypic variance, 461
  mapping, 504–507
  mean phenotype, 454
  Mendelian markers for, 504–507
  mutation-selection balance, 501–504
  number of genes affecting, 485–489
  phenotypic correlation, 490
  random genetic drift and, 498–501
  scaling, 487
  stabilizing selection, 501
  threshold, 432, 481–485
Quantitative trait loci, 504–507

r-selection, 523, 537
Rabbit, 346, 389
Race, 301–306
Racial admixture, 308–309
Radiation, effect on mutation, 103–107, 108–109
Random genetic drift
  adaptive topography and, 323–326
  allele frequency and, 71–75, 80–82
  defined, 61
  diffusion approximation and, 71–74
  fitness and, 88–90, 206–209
  founder effect, 84–85
  inbreeding and, 74–82
  infinite alleles model and, 122–126, 130
  infinite sites model, 132–135
  migration and, 306–320
  Moran model, 87–88, 377–378
  natural selection and, 206–209
  pure, 206–207
  quantitative traits and, 498–501

shifting balance theory and, 323–326
in subpopulations, 306–313
in wildlife management, 86–87
Wright-Fisher model of, 61–63, 66–67
see also Neutral theory of molecular evolution
Random mating, 31, 34–35
two loci, 45–54
X-linked genes and, 43–45
Rare alleles, 19, 39, 45, 142
Rat, Norway, 401, 492
warfarin resistance, 198–199
Rate of evolution, 356, 361–362
Recessive allele, 38–39
favored, 155–156
inbreeding depression, 253–255
lethals, 199–200
rare, 39
see also Dominance
Recombination, 28, 338, 382
intrachromosomal, 114
intragenic, 133–134
inversions and, 99
linkage disequilibrium and, 47, 130–132, 267–268
linkage equilibrium and, 46
suppression of, 193
viability selection, 211–212
Recombination frequency, 47–50
see also Linkage disequilibrium
Recurrent mutation, 19, 199, 201
Recurrent sib mating, 258–259
Red Queen hypothesis, 515
Regression, 331, 362, 435–438, 468
Regulatory genes, 24
Relaxed selection, 446–447
Repetitive DNA
dispersed, *see* Dispersed repetitive DNA
Drosophila repeated elements, 413–416
human *Alu* sequences, 346, 402–406
localized, *see* Localized repetitive DNA
rabbit C repeats, 346
rate of disappearance of, 405–406
SINEs and LINEs, 402–406
Reproductive isolation, 588–590
Reproductive value, 535
Resistance transfer factor, 413, 580–586
Restriction enzyme, 331, 333
Restriction fragment length polymorphism (RFLP), 13–14, 26–27, 140
and β-globin, 339
Restriction map variation, 335–336
Restriction site, 11–13, 333
haplotypes, 26
mapping, 26
Mus domesticus (mouse), 29
silent sites, 17
Restriction site variation, 13, 25–26, 92, 330–339
estimating heterozygosity 331–335
estimating species divergence 334
Retrovirus, 404, 416–420
Reverse transcriptase, 404
RFLP, *see* Restriction fragment length polymorphism
Rh blood group, 38–39, 186
Rhagoletis (apple maggot flies), 596
Ribosomal proteins, 24, 339
Ribosomal RNA genes
and abnormal abdomen, 547
in *Drosophila*, 393–395
5S, 330
Ribulose-1,5–bisphosphate carboxylase, 392–393
RNA polymerase, 550
RNA processing, 339
RNA replicase, 365
Robertsonian translocation, 100–101

Salmonella typhimurium, 114, 358, 370, 585, 585
Salmonidae (salmon, trout and char), 603
Sampling
binomial, 66, 207
error, 65
random genetic drift and, 70
with replacement, 62–63
variance, 87
Sanger dideoxy sequencing, 14–15, 341–342
Satellite DNA, 399–402
Scaling effects and quantitative traits, 487
Scallop, 421
Schizophrenia, 432
Segregation, Mendel's law of, 1–2
Segregation distorter, 192–193
Selection
artificial, *see* Artificial selection
differential, 443
index, 489
natural, *see* Natural selection
response, 442–450
Selective constraints on nucleotide substitution, 371, 375
Selective neutrality, *see* Neutral theory of molecular evolution
Self-fertilization, 235, 255–261
fixation index, 262
genotype frequencies, 244–245
inbreeding coefficient, 80
linkage disequilibrium, 263–268
mixed mating model, 261–263
recurrent, 257
see also Inbreeding
Self-incompatibility, 274–277
Selfing, *see* Self-fertilization
Selfish DNA, 399, 403, 406
Senescence, 535–538
experimental studies of 537–538
Sequential electrophoresis, 23–24
Sex determination, 558–560
Sex ratio, 86–87, 555–560
Drosophila pseudoobscura SR chromosome, 557–559
environmental sex determination, 559
Fisher's rule, 555
local mate competition 558
modifier theory, 557
population genetic theory, 556–557
Sexual asymmetry, 191
Sexual selection, 148, 194–198
coefficient of, 196–197
Fisher's runaway process, 194–195
Shark, 359, 398
Sheep, 194, 470
Sheep blowfly (*Lucilia cuprina*), 474
Shifting balance theory, 323–326, 515
Shigella, 583
Sib analysis, 476–477, 537
Sib mating, 255, 258–260
Sibling species, 588
Sickle cell anemia, 7, 99, 162, 169–170, 198, 284
Signal peptide, 375
Silent substitution (synonymous substitution), 17, 97, 103, 343
SINEs
human *Alu* repeats, 346, 402–406

rabbit C repeats, 402
Six-cutters, 11
Skin graft acceptance in cheetahs, 23
Skua, arctic (*Stercorarius parasiticus*), 197–198
Snowshoe hare (*Lepus americanus*), 85
Sociobiology, 567
Soft selection (Levene model), 525–528
SOS repair in *E. coli*, 101–103
Southern blotting, 12–13, 26–28, 140, 330, 334, 336, 392
Sparrows, 497
Speciation, 586–597
  allopatric 591–595
  parapatric 595–596
  sympatric 596–597
Species genetic differences, 597–604
Sperm displacement, 193
Sporophytic and gametophytic selection, 204–206
Sporophytic self-incompatibility, 274–275
SR chromosome, 557–559
Stability, of equilibria, 159–161
Stabilizing selection, 380, 381, 501, 504
Stable age distribution, 533
Standard deviation, 433–434
Start codon, 98–99
Stationary distribution, 69, 208
Statistical hypothesis testing
  binomial distribution, 62, 66–67
  Ewens-Watterson test, 130, 140–143, 376
  null hypothesis, 38, 61, 116, 138–139
  power, 55
  sampling, 61–63
Statistics, 436–438
  central limit theorem, 433–435
Steady state
  effective population size and, 87
  infinite alleles model and, 122–130
  Markov chains and, 68
  probability, 133–134
Stepping stone model, 317–319, 322
Stepwise mutation model, 135–136
Stirling numbers, 128
Stop codon, 98–99
Streptomycin resistance, 409, 582
Subpopulation(s), 281–293
  estimating migration among, 319–323
  extinction and recolonization, 316
  genetic distance, 301–306
  genetic divergence among, 293–301
  interdeme selection, 323–329
  island model and, 306–316
  isolated, 282
  random genetic drift and, 74–82, 297, 306–313
  semi-isolated (demes), 281
  stepping stone models, 317–319
  Wahlund effect, 282–293
Suicide, adaptive, 563
Supergene, 52, 99, 162
Survivorship curve, 530–531
Swallow (*Hirundo rustica*), 197–198
Swordtail (*Xiphophorus*), 558
Sympatric speciation, 596–597
Synonymous substitution (silent substitution) 97, 103

t alleles, 191–193, 201
Taricha rivularis (salamander), 138
Tay Sachs disease, 284, 324–325
Taylor approximation, 159–161
Temporal isolation and speciation, 597
Temporal variation in selection, 529–530
Test statistic, 38
Tetrahedron, 46–47
Thalassemia, 339, 391, 403
Thermal stability, 329
Thermus aquaticus, 344
Threshold traits, 432, 481–485
Thrum, 276
Thymine dimers, 101–103
Tigriopus californicus (copepod), 535, 544, 547
Tissue plasminogen activator (TPA), 397
TIT FOR TAT, 568–569
Tobacco, 393
Tomato (*Lycopersicon*), 393, 488, 506
Trade-off (*r* vs. *K* selection), 523
Trans (repulsion phase), 209
Transfer RNA, 24
Transition mutation, 97, 116, 118
Transition probability, 67–69
  binomial, 207–208
  coalescence, 91–92
  matrix, 68–69
  mutation and selection and, 69
  see also Wright-Fisher model
Translation, 339
Translocations, 100–101
Transposable elements, 27, 407–416
  attributes of, 407–409
  in bacteria, 409–413
  copia, 413–415
  in *Drosophila melanogaster*, 27, 413–416
  in eukaryotes, 413–415, 416–420
  induced mutations from, 403
  mariner, 414
  mutations and, 96, 403
  population genetics theory of, 407–408, 410
  roo, 408
  transposons, 411–413
Transposase, 407
Transposition mutations, 96
Transposons, 411–413, 584
  *Tn*10, 411–412
Transversions, 116
Triangle inequality, 174
Tribolium (flour beetle), 435–436, 447, 488, 562
Triticum vulgare (wheat), 257, 450–454, 485
Truncation selection, 442–446
Turnip, 424
Twin studies, 480–481, 484–485

Underdominance, 154, 163–165, 167
Underdominance in carrying capacity, 521
Unequal crossing over, 384–393
UPGMA tree, 378
UV radiation, 101

Vagility, 593
Variance, 433–436
Variance–covariance matrix, 494–496
Variance effective size, 82, 87
Variation, intraspecific
  genetic, 4–7
  hidden, 7–8, 199
  multifactorial (quantitative), 430–435
  phenotypic diversity and, 6
  restriction site, 25–26, 330–339

Viability
frequency dependent, 184–186
fundamental theorem, 164–168
haploid, 182–183
multilocus, 217–220
one locus many alleles, 168–177
one locus two alleles, 151–164
two loci, 209–217
two sexes, 177–180
X linkage, 180–182
Viceroy, 572–574
Virogenes, 416, 418–419

Wahlund's principle, 282–296, 301
Wallace effect, 591
Warfarin and resistant rats, 198–199
Warning coloration, 570
Whale, 401–402, 602–603
Wheat (*Triticum vulgare*), 257, 450–454, 485
Wildlife management, 86–89
Wild oats (*Avena barbata*), 260
Wright-Fisher model
diffusion approximation and, 73–74
effective population size, 88–90
Markov chains and, 67–70
natural selection and, 207–209
random genetic drift and, 66–67
viability and, 206–209

X chromosome, 43, 243–244
sex ratio (SR), 557–559
speciation and the, 600–601
transposable elements on, 408
X-linked genes
Hardy-Weinberg principle and, 43–45
heterozygosity and, 79
inbreeding coefficient and, 243–244
linkage disequilibrium and, 55
mutation rate, 103–105
rDNA in *Drosophila*, 393–395
viability selection, 180–182
Xanthine dehydrogenase (XDH), 24, 126, 141–143
Xeroderma pigmentosum, 101

Y chromosome, 92
rDNA in *Drosophila*, 393–395
speciation and the, 600–601
Yanomama Indians, 300–303
Yeast, 27

Zoarces viviparus (eelpouts), 138, 224–227
Zoos, 89
Zygote, 148, 211
Zygote elimination, 275–277

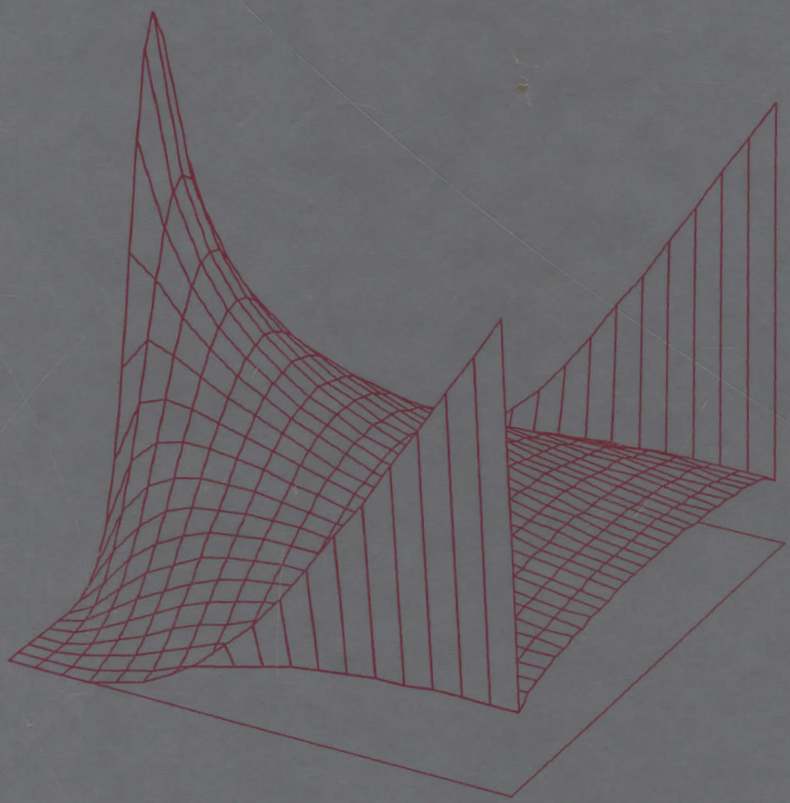
ABOUT THE BOOK

The typefaces used in this book are Palatino and Helvetica, both of which were designed by the contemporary German calligrapher and type designer, Hermann Zapf. The type was generated at DEKR Corporation on their Linotron 202.

Joseph J. Vesely created the book format and coordinated production.

The cover design is by Rodelinde Graphic Design; covers were printed by New England Book Components.

This book was manufactured by R. R. Donnelley & Sons.



ISBN: 0-87893-302-6